Aprender

SERIE DE DOCUMENTOS TÉCNICOS / 4

APRENDER 2018 CONSTRUCCIÓN DE LAS PRUEBAS:

CONTENIDO Y PROPIEDADES PSICOMÉTRICAS

Secretaría de Evaluación Educativa

Ministerio de Educación, Cultura, Ciencia y Tecnología
Presidencia de la Nación

SERIE DE DOCUMENTOS TÉCNICOS / 4

APRENDER 2018

CONSTRUCCIÓN DE LAS PRUEBAS:

CONTENIDO Y PROPIEDADES PSICOMÉTRICAS

AUTORIDADES

Presidente
Mauricio Macri

Jefe de Gabinete de Ministros
Marcos Peña

Ministro de Educación, Cultura, Ciencia y Tecnología de la Nación
Alejandro Finocchiaro

Secretario de Gobierno de Cultura
Pablo Avelluto

Secretario de Gobierno de Ciencia, Tecnología e Innovación Productiva
Lino Barañao

Titular de la Unidad de Coordinación General del Ministerio de Educación, Cultura, Ciencia y Tecnología
Manuel Vidal

Secretaria de Evaluación Educativa
Elena Duro

Secretario de Gestión Educativa
Oscar Ghillione

Secretario de Políticas Universitarias
Pablo Domenichini

Secretaria de Innovación y Calidad Educativa
Mercedes Miguel

COORDINACIÓN

Prof. Elena Duro

EQUIPO A CARGO DE LA ELABORACIÓN DEL DOCUMENTO

María Aranguren

María Elena Brenlla

Augusto Hoszowski

Agnés Laxague

Quimey Lassus

Elisa Marilina Zayas

ÍNDICE

INTRODUCCIÓN

El propósito de las pruebas Aprender 2018 es continuar y fortalecer el compromiso asumido con las escuelas desde los primeros operativos nacionales de evaluación hasta hoy día.

Una evaluación educativa nos brinda información clave sobre los logros y resultados de los estudiantes. Es una foto panorámica o un corte transversal que muestra cuáles son los aprendizajes de los estudiantes en un determinado momento. Sin embargo, una evaluación también es un proceso ya que implica la recolección sistemática de información y una interpretación de los resultados en función de ciertos criterios que permiten realizar un diagnóstico preciso del sistema educativo y que colabora a la toma de decisiones para acciones orientadas a las mejoras del sistema.

En el mes de octubre de 2018, se aplicaron las pruebas Aprender 2018 en las áreas de Matemática y Lengua en 6° año de la Educación Primaria.

Dichos instrumentos fueron diseñados siguiendo los parámetros y condiciones estipulados para obtener resultados válidos y confiables. Para ello, se tuvieron en cuenta los lineamientos de los *Estándares para la Evaluación Educativa y Psicológica* (American Educational Research Association [AERA], American Psychological Association [APA] & National Council on Measurement in Education [NCME], 2014). En este documento se detalla cómo ha sido la construcción de cada una de las pruebas y se aportan evidencias acerca de su validez y fiabilidad. Primero se define qué son las pruebas referidas a criterio y su diferencia con las referidas a normas, luego se señalan los requisitos psicométricos de fiabilidad y validez que debe reunir una medida y, por último, los pasos seguidos para el diseño de las pruebas Aprender 2018.

PRUEBAS CRITERIALES Y PRUEBAS NORMATIVAS

En consonancia con las evaluaciones nacionales previas, las evaluaciones Aprender 2018 mantienen el mismo enfoque pedagógico que el utilizado en ocasiones anteriores. En este sentido, al igual que en Aprender 2016 y 2017 y que en los Operativos Nacionales de Evaluación (ONE) 2007, 2010 y 2013, se busca conocer los contenidos, destrezas y capacidades que los estudiantes logran dominar en cada una de las áreas y años evaluados. Este tipo de pruebas se denominan pruebas referidas a criterio y se diferencian de las pruebas referidas a la norma (Glaser, 1963).

En apariencia, las pruebas criteriales y las pruebas normativas no muestran grandes diferencias ya que ambas evaluaciones suelen estar conformadas por ítems o ejercicios que siguen un formato similar y que requieren que el estudiante ponga en juego las mismas capacidades cognitivas en resolverlos (Leones, 2005). Sin embargo, existen diferencias sustanciales en la construcción y en las inferencias que es posible hacer a partir de cada una de ellas.

Una prueba normativa busca identificar cuál es la posición de un estudiante en relación con los demás estudiantes evaluados, mientras que una evaluación criterial, busca conocer el dominio que tiene un alumno de ciertos criterios preestablecidos. Por ejemplo, en términos descriptivos, el desempeño de un estudiante en una prueba referida a criterio podría indicarse como "Enrique ha respondido correctamente al 75% de las preguntas de la prueba de Lengua" que difiere de "El desempeño de Enrique supera al 75% de los estudiantes que respondieron a la prueba de Lengua", propio de las pruebas referidas a normas.

A partir de los cambios introducidos en la educación en las décadas pasadas -en los que se dio especial énfasis a los programas e intervenciones educativas-, fue necesario revisar las evaluaciones referidas a

las normas de grupo ya que no resultaban apropiadas para conocer los rendimientos educativos ni para diagnosticar necesidades de intervención (Martínez Arias, 1996). Tal como lo ilustra el ejemplo anterior, ¿de qué sirve conocer la posición de un estudiante en su clase si no sabemos si logró alcanzar los objetivos básicos de la materia? Enrique podría superar al 75% de sus compañeros contestando correctamente unos pocos ítems más que ellos y aun así no haber logrado los objetivos propuestos en el programa o en la evaluación (por ejemplo, contestar correctamente al 50% de los ítems de la prueba).

En relación con lo anterior, cabe aclarar que las pruebas criteriales han sido desarrolladas con el objetivo de corregir aquellas inadecuaciones de las pruebas normativas, particularmente, sus limitaciones para ser utilizadas en evaluaciones estandarizadas (Jornet Meliá & González Such, 2009). Así, las pruebas criteriales permiten interpretar los resultados obtenidos por los estudiantes de acuerdo a ciertos criterios absolutos referidos al dominio que se está evaluando.

Las pruebas Aprender 2018 siguen la línea inaugurada en 2003 en nuestro país y se realizan bajo el enfoque de pruebas referidas al criterio. Ello implica que los dominios –sean objetivos, capacidades, destrezas o competencias- están bien definidos, que los ítems los cubren apropiadamente y que si se establecen estándares de rendimiento (o puntos de corte) esto se realizará con la metodología apropiada. Las pruebas criteriales se conforman por ítems o reactivos que responden a los objetivos del aprendizaje. Como se indicó, los ítems incluidos en la prueba son un conjunto representativo de un dominio claramente definido (Hambleton & Swaminathan, 1978), permiten conocer la ubicación de un sujeto en un continuo representativo del dominio evaluado –y no respecto del grupo normativo- y se obtiene una interpretación directa de la puntuación obtenida: el desempeño que exhibe el estudiante indica su grado de competencia (Leyva Barajas, 2011).

EVIDENCIAS DE VALIDEZ Y FIABILIDAD DE UNA PRUEBA EDUCATIVA

En la construcción de una prueba educativa –al igual que en cualquier construcción de un test– se buscará que sus resultados sean válidos y confiables. Para ello será necesario contar con evidencias suficientes que demuestren que el instrumento utilizado funciona correctamente. En ocasiones es posible realizar algunos ensayos o pruebas piloto[1] para corregir posibles errores en el instrumento y evitar errores en la medida.

El concepto de validez se utiliza para indicar el grado con el que un instrumento efectivamente mide aquello que se propone o debe medir (Martínez Arias, 1995). La fiabilidad, por su parte, es una propiedad psicométrica que hace referencia a la precisión de la medida (Cronbach, 1972). Un test es confiable cuando al repetirse su aplicación –en un período de tiempo relativamente acotado y en un mismo grupo de participantes– arroja resultados similares a los obtenidos en ocasiones previas (Cortada de Kohan, Macbeth & Alonso, 2008). A su vez, una prueba es válida cuando las inferencias realizadas a partir de ella son adecuadas, significativas y útiles (AERA, APA & NCME, 2014).

Actualmente, se sostiene que la validez y la fiabilidad no son atributos fijos o inmutables de un instrumento, sino que se trata de una serie de juicios que aportan evidencias acerca de qué tan apropiadas son las inferencias que puedan realizarse de una prueba en un determinado contexto o para un determinado

1 Prueba administrada a una muestra representativa de la población a fin de obtener resultados preliminares del funcionamiento de los ítems que conforman un instrumento. Los resultados son utilizados luego para realizar ajustes y conformar la versión final de la prueba.

grupo (AERA, APA & NCME, 2014, Covacevich, 2014). Esto se debe a que los estudios de validez y fiabilidad de un instrumento se efectúan sobre muestras delimitadas, con características propias, y respuestas y rendimientos concretos. Un instrumento puede ser válido para una población y no para otra. Por ejemplo, una prueba de razonamiento lógico diseñada para alumnos de 5° año no podrá reflejar de manera válida las habilidades de razonamiento lógico de un estudiante de 2° año. De ahí que en el área de la psicometría y de la investigación educativa, se hable entonces de evidencias de validez o evidencias de fiabilidad para un determinado grupo o determinado contexto.

Evidencias de validez

Teniendo en cuenta que la validez está relacionada con las inferencias que podrán realizarse a partir de los resultados de una prueba, para juzgar su calidad se deberán reunir varios indicadores y tipos de información (Covacevich, 2014). Estos indicadores y evidencias pueden ser obtenidos a través de distintas fuentes y enfoques. De acuerdo a las fuentes y/o enfoque utilizado, tendremos evidencias de distintos tipos de validez. Tradicionalmente, se suele hablar de evidencias referidas a la validez de contenido, referidas a la validez de criterio y evidencias referidas a la validez de constructo (Leyva Barajas, 2011).

Las evidencias de validez de las pruebas criteriales deberán ser analizadas en función del propósito para el cual son diseñadas. Básicamente, una prueba criterial busca describir el dominio que tienen los participantes de un área disciplinar y clasificarlos según niveles de desempeño de acuerdo a sus habilidades y capacidades (Hambleton, Swaminathan & Rogers, 1991). De ahí que se buscará, en primer lugar, reunir evidencias de la validez de contenido de la prueba.

El contenido de una prueba educativa estará delimitado por los contenidos y capacidades evaluadas de acuerdo a la disciplina y año. Por otra parte, las evidencias de validez de contenido pueden ser obtenidas en diferentes etapas: (a) proceso de construcción de ítems; (b) proceso de puesta prueba de los ítems y; (c) proceso de selección de los ítems que conformaran la versión final de la prueba. Cada etapa se caracteriza por presentar diversas fuentes de información.

En la primera etapa, se deberá conformar un grupo de itemistas y de supervisores (e.g. lectores críticos, expertos en psicometría y expertos del Instituto Nacional de Formación Docente y de la Dirección de Innovación Educativa que atiendan a las tablas de especificaciones en donde se delimitan los contenidos y capacidades a evaluar para la confección, evaluación y edición de los reactivos. En segundo momento, se tendrán en cuenta los resultados y análisis de la prueba piloto. Por último, se realizará una selección y ajuste de los ítems para la prueba final. Cada una de estas etapas, brinda información acerca de la validez de contenido de la prueba y también acerca de su fiabilidad. En el próximo apartado se indican los procedimientos llevados a cabo para analizar la validez de contenido de las pruebas Aprender 2018.

Evidencias de fiabilidad

En las pruebas criteriales, la confiablidad es analizada a partir de la Teoría de Respuesta al Ítem (TRI) o Teoría del Rasgo Latente. Uno de los supuestos de la TRI es que existe una relación directa entre los resultados obtenidos en un ítem o en un conjunto de ítems, y el grado en que se posee la variable o rasgo evaluado (Hambleton et al, 1991; Leones, 2005, Secretaría de Evaluación Educativa [SEE], 2017). Es decir, que el acierto en un ítem o en un conjunto de ítems, implica el dominio de los contenidos y capacidades evaluadas. Así, el postulado principal de los modelos TRI es que existe una relación funcional entre el rasgo latente que miden los ítems y la probabilidad de acertar a éstos (Muñiz, 2010). Esta relación funcional se llama Curva Característica del Item (CCI). Hay una variedad de formas posibles funcionales y, por lo tanto, de modelos

TRI en la literatura psicométrica (e.g. modelo logístico de un parámetro o modelo de Rasch, modelo de dos parámetros, modelo de tres parámetros).

En consonancia con la mayoría de las evaluaciones internacionales, las propias ONE desde 2005 y Aprender 2016 y 2017, las pruebas Aprender 2018 han sido analizadas con base en la TRI. En la prueba Aprender 2016, 2017 y 2018 se optó por el modelo logístico de dos parámetros y en el piloto de Aprender 2018 (como en los pilotos de años anteriores) por la logística de un parámetro.

El uso de la TRI permite observar qué tan bien un ítem o un conjunto de ítems, discrimina entre estudiantes que tienen un alto y un bajo desempeño en las áreas evaluadas. Por otra parte, también se obtienen otros indicadores como el índice de dificultad y el índice de correlación biserial. Además, facilita una comparación más precisa de los resultados a través del tiempo (SEE, 2017).

Por último, dado que las pruebas criteriales tienen por objeto medir el dominio que tiene un estudiante de una determinada área o disciplina, la representatividad y relevancia de los ítems incluidos es parte fundamental del proceso. De ahí que obtener evidencias acerca de la validez de contenido de las pruebas sea un requisito clave para garantizar idoneidad de los resultados obtenidos.

Otros aspectos no menos relevantes que hacen a la fiabilidad de la prueba son los referidos a las condiciones estandarizadas de aplicación de los instrumentos. Si las condiciones de aplicación no son similares entre los diferentes grupos, los resultados no serán comparables. En este sentido, se deben considerar los tiempos de aplicación de la prueba, el entrenamiento de los aplicadores (la estandarización de las instrucciones para todos los participantes), y la facilidad para la puntuación y análisis de las respuestas obtenidas. Mientras más cuidados y sistematizados sean estos elementos, más reguardada estará la comparabilidad de los resultados obtenidos por distintos grupos de participantes.

En el próximo punto se indican las fases en la construcción de una prueba referida a criterio y qué procedimientos concretos se llevaron a cabo para el análisis de las evidencias de fiabilidad y validez de las pruebas Aprender 2018.

CONSTRUCCIÓN DE PRUEBAS REFERIDAS A CRITERIO
LAS PRUEBAS APRENDER 2018.

Existen criterios consensuados por la comunidad científica para el diseño de una prueba referida a criterios. En concreto, Hambleton et al. (1991) propusieron etapas que se resumen a continuación a la vez que se indican qué procedimientos concretos se llevaron a cabo en la construcción de las pruebas Aprender 2018.

1) Planteamientos iniciales

En esta etapa se plantean las hipótesis acerca de los propósitos y objetivos de las pruebas, así como a qué grupos está dirigida, cómo se planificarán los tiempos y qué recursos humanos y materiales se requieren, a quiénes se convocará como expertos y cuál será la longitud y el tiempo para cada prueba.

Tabla 1. *Propósito y caracterización de las pruebas Aprender 2018*

Nombre	Pruebas Aprender 2018 Lengua y Matemática
Propósito y objetivos	Evaluar capacidades y contenidos que debería alcanzar un alumno al fin del nivel primario de acuerdo a los NAP
Grupos	Estudiantes regulares de 6° Nivel Primario
Cobertura	Censal
Tipo de ítems	Cerrados de opción múltiple
Planificación	Piloto Diseño gráfico
Identificar Expertos	Cuerpo colegiado de lectores críticos expertos de todo el país. Lectores críticos disciplinares internos. Expertos del Instituto Nacional de Formación Docente (INFoD) y de la Dirección de Innovación Educativa
Longitud y tiempo de aplicación	30 ítems 60 minutos por prueba (más 10 minutos extra)

2) Revisión de los objetivos

En esta etapa se revisan críticamente los objetivos de evaluación, se selecciona el conjunto de objetivos, capacidades y contenidos que evaluarán las pruebas y se preparan las tablas de especificaciones de cada una.

Siguiendo estos lineamientos y con los antecedentes nacionales, las pruebas Aprender 2018 fueron construidas a partir de los Núcleos de Aprendizaje Prioritarios (NAP), los diseños curriculares jurisdiccionales y los consensos jurisdiccionales. Posteriormente y, al igual que en Aprender 2016 y 2017, en función del marco general de referencia, se elaboraron las estructuras de cada una de las pruebas a partir del diseño de la tabla de especificaciones en la que se indican los contenidos y las capacidades a evaluar, así como el valor relativo a cada una de ellas.

La consecución de los objetivos para las pruebas Aprender 2018 estuvo a cargo del equipo pedagógico de la SEE quienes, en consulta e intercambio continuo con especialistas de otras áreas del Ministerio de Educación, Cultura, Ciencia y Tecnología y del campo educativo en general, realizaron las definiciones de capacidades y contenidos y establecieron los fundamentos teóricos sobre las que se basó la evaluación en cada disciplina.

Cabe destacar que el equipo de la SEE se encuentra conformado por docentes expertos –y actualmente, en ejercicio– en cada una de las áreas disciplinares principales (Matemática, Lengua, Ciencias Naturales y Ciencias Sociales).

En Aprender se evalúan distintas capacidades y contenidos. Se entiende por contenidos a los saberes relevantes que los alumnos y las alumnas que concurren a las escuelas deben aprender, y que los maestros deben enseñar. Se entiende por capacidades cognitivas a aquellas operaciones mentales que el sujeto utiliza para establecer relaciones con y entre los objetos, situaciones y fenómenos. Se logran a través del proceso de enseñanza y del proceso de aprendizaje y cobran significado de acuerdo con la determinación de contenidos socialmente relevantes y altamente significativos, frente a los que se ponen en juego y a través de los cuales se desarrollan (Ministerio de Educación, Ciencia y Tencología - DiNIECE, 2001)

A continuación, se muestran las capacidades y contenidos delimitados para cada área disciplinar y año considerados para la elaboración de los ítems de las pruebas piloto Aprender 2018.

Evaluaciones en 6º grado de nivel primario

MATEMÁTICA

Capacidades cognitivas:

- RECONOCIMIENTO DE DATOS Y CONCEPTOS
  Capacidad cognitiva de identificar datos, hechos, conceptos, relaciones y propiedades matemáticas, expresados de manera directa y explícita en el enunciado.

- RESOLUCIÓN DE SITUACIONES EN CONTEXTOS INTRAMATEMÁTICOS Y/O DE LA VIDA COTIDIANA
  Capacidad cognitiva de solucionar situaciones problemáticas contextualizadas, presentadas en contextos que van desde los intramatemáticos hasta los de la realidad cotidiana.

- COMUNICACIÓN EN MATEMÁTICA
  Interpretar información: comprender enunciados, cuadros, gráficos; diferenciar datos de incógnitas; interpretar símbolos, consignas, informaciones; manejar el vocabulario de la Matemática; traducir de una forma de representación a otra, de un tipo de lenguaje a otro.

- SOLUCIÓN DE OPERACIONES
  Resolver operaciones en los distintos conjuntos numéricos usando distintos procedimientos.

Tabla 2. *Tabla de contenidos Matemática 6° año de la Educación Primaria*

Bloque	Contenidos
Números y operaciones	• Números naturales, fraccionarios y expresiones decimales. Reconocimiento y uso. • Sistema decimal de numeración. Características. • Representación y ubicación de naturales, fraccionarios y decimales en la recta numérica. • Operaciones: suma, resta, multiplicación y división entre naturales, decimales y fraccionarios (excluida la división entre decimales y entre fraccionarios) • Resolución de problemas que requieran diferentes significados de las cuatro operaciones, incluida la proporcionalidad con constante entera.• Relaciones entre números: divisibilidad.• Resolución fundamentada de cálculos y/o situaciones problemáticas
Geometría y medida	• Relación entre sistemas de unidades: longitud, capacidad, peso, superficie y tiempo. • Cálculo de medidas: estimación. Aproximación y exactitud.• Perímetro: concepto. Perímetro de polígonos regulares. • Área: concepto. Unidades. Equivalencias. Área de polígonos comunes.• Sistemas de referencia para la ubicación de puntos en un plano. • Figuras geométricas: reconocimiento de elementos y propiedades de triángulos, cuadriláteros, circunferencia y círculo. • Cuerpos geométricos: reconocimiento y propiedades de prismas, pirámides, cubo, cilindro, cono y esfera. • Resolución de problemas que requieran analizar, describir, comparar, clasificar y construir figuras en base a las propiedades conocidas.
Estadística y probabilidad	• Expresión e interpretación de datos a través de cuadros, diagramas y gráficos estadísticos. • Resolución de problemas que requieran interpretación de datos explícitos e implícitos en diferentes gráficos.

LENGUA

Capacidades cognitivas

- EXTRAER:
  Localizar información en una o más partes de un texto. Los lectores deben revisar, buscar, localizar y seleccionar la información. Deben cotejar la información proporcionada en la pregunta con información literal o similar en el texto y utilizarla para encontrar la nueva información solicitada.

- INTERPRETAR:
  Reconstruir el significado global y local; hacer inferencias desde una o más partes de un texto. Los lectores deben identificar, comparar, contrastar, integrar información con el propósito de construir el significado del texto.

- REFLEXIONAR Y EVALUAR:
  Relacionar un texto con su propia experiencia, conocimientos e ideas. Los lectores deben distanciarse del texto y considerarlo objetivamente. Deben utilizar conocimiento extra-textual (la propia experiencia, elementos proporcionados por la pregunta, conocimiento del mundo, conocimiento de la lengua, conocimiento de distintos géneros discursivos). Los lectores deben justificar su propio punto de vista.

Los contenidos evaluados en Lengua corresponden a cada una de las capacidades cognitivas para 6° año de Nivel Primario.

Tipos de textos evaluados: cuentos breves de autores consagrados y textos expositivos o argumentativos tales como columnas de opinión, ensayos breves y textos académicos o de divulgación científica provenientes de revistas especializadas o manuales

Tabla 3. *Tabla criterial Lengua 6° año de la Educación Primaria*

Bloque	Contenidos
EXTRAER	• Información explícita en texto literario y no literario. • Secuencia en texto literario. • Resumen.
INTERPRETAR	• Tema en texto literario y no literario. • Relaciones textuales. • Procedimientos de cohesión. • Características de personajes. • Vocabulario. • Información inferencial. • Relación texto-paratexto.
REFLEXIONAR Y EVALUAR	• Recursos literarios. • Tipos de narradores. • Tipologías y géneros discursivos

3) Redacción de los ítems

En esta fase se construyen y editan ítems con el objetivo de probarlos en el estudio piloto posterior. Los ítems incluidos en las pruebas de criterio pueden ser de respuesta abierta o ítems de respuesta cerrada y, en ambos casos, la cuestión esencial es delimitar con claridad el dominio y el criterio sobre la que se basa la prueba. Los ítems de respuesta abierta requieren que el estudiante construya o elabore la respuesta, mientras que los ítems de respuestas cerrada obligan a la selección de una o varias opciones de entre las propuestas.

Las evaluaciones educativas de gran escala suelen utilizar ítems de respuesta cerrada dado que su aplicación y corrección es más sencilla y más rápida (Covacevich, 2014). Además, las evaluaciones que utilizan ítems de opción múltiple pueden ser corregidas mediante el uso de lectores ópticos, lo que no solo agiliza los procesos de corrección, sino que también implica menores costos y evita las diferencias de criterio que podrían existir entre diferentes correctores para los ítems de respuestas abiertas (Livingston, 1999).

En el caso de las pruebas Aprender 2018 se utilizaron ítems de elección múltiple para las evaluaciones de Matemática y Lengua en 6° año de la Educación Primaria. Estos ítems están compuestos por un enunciado y cuatro opciones de respuesta predeterminadas, de las cuales solo una es la respuesta correcta (ver tabla 4).

Tabla 4. *Tipos de ítems en las evaluaciones Aprender 2018*

Nombre	Año y Nivel	Tipo de ítems
Aprender 2018 Matemática	6° Nivel Primario	Ítems cerrados, opción múltiple
Aprender 2018 Lengua	6° Nivel Primario	Ítems cerrados, opción múltiple

Para la elaboración y redacción de los ítems se conformó un cuerpo colegiado de constructores de ítems. La convocatoria para la selección de los perfiles se realizó a través de las Unidades de Evaluación Jurisdiccional (UEJ) a fin de garantizar la representatividad de todas las provincias. Los requerimientos para participar fueron los siguientes: (a) docentes con título profesional en el nivel primario; (b) ejercicio actual de la profesión en 6° año del nivel primario en las áreas de lengua y matemática; (c) experiencia de al menos dos años en el ejercicio de la docencia preferentemente en escuelas de gestión pública y privada; (d) capacidad de trabajar en forma autónoma en base a un cronograma y; (e) disponibilidad para capacitarse. Se valoraron aquellos perfiles que presentaban especializaciones y/o experiencia profesional en evaluación de aprendizajes.

El cuerpo colegiado de ítemistas se conformó por 21 profesionales del área de Matemática pertenecientes a las provincias de: Ciudad Autónoma de Buenos Aires, Provincia de Buenos Aires, Catamarca, Córdoba, Entre Ríos, Jujuy, Mendoza, Neuquén, San Luis, Tucumán y Tierra del Fuego y 10 profesionales del área de Lengua pertenecientes a las provincias de: Ciudad Autónoma de Buenos Aires, Corrientes, Mendoza, Santiago del Estero y Tucumán.

Todos los ítemistas seleccionados recibieron una *Guía para la elaboración de ítems de opción múltiple* (en dicha guía se encontraba la tabla especificaciones de cada área y año), un video tutorial elaborado por el equipo de especialistas en psicometría de la SEE y, un documento con pautas para la selección de las imágenes adjuntas al ítem construido. Cada ítemista debía completar un formulario con los datos cada ítem construido especificando las características del mismo. Se indicó que los ítems redactados debían cubrir el continuo de los niveles de dificultad de la prueba: Nivel Por debajo del nivel Básico, Nivel Básico, Nivel Satisfactorio y Nivel Avanzado. La figura 1 ilustra la secuencia que se siguió para la elaboración de los ítems.

Cada profesional contratado enviara 50 ítems del año y área evaluado. Los especialistas del equipo pedagógico de la SEE realizaron un seguimiento de las tareas realizadas. La figura 1 ilustra la secuencia que se siguió para la elaboración de los ítems.

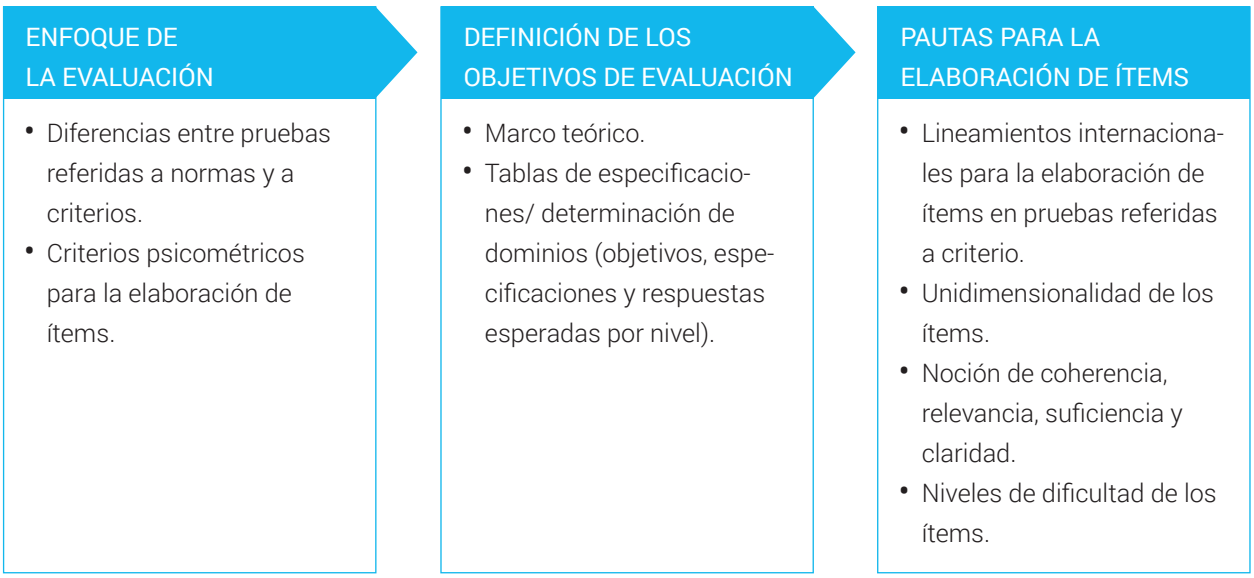

Figura 1. Secuencia en la capacitación de la elaboración de los ítems.

4) Evaluación de la validez de contenido

Tal como se señaló, en una prueba referida a criterio es crucial el análisis de las evidencias de validez de contenido. Ello implica la identificación de un grupo de expertos en el contenido de las pruebas y en psicometría, el examen de la concordancia entre ítems, contenidos y objetivos de evaluación, la adecuación técnica y la revisión global de los ítems. A continuación, se indican los procedimientos seguidos para el análisis de las evidencias de validez de contenido de las pruebas Aprender 2018. Cabe indicar que todos los actores involucrados en cada una de las etapas de construcción de la prueba firmaron acuerdos de confidencialidad a fin de garantizar la fiabilidad de la prueba.

Evaluaciones de Matemática y Lengua

Una vez elaborados los ítems, se procedió a la revisión por parte de: (a) especialistas de las áreas (equipo pedagógico de la SEE); (b) expertos en psicometría; (c) cuerpo colegiado de lectores críticos y; (d) profesionales del INFoD y de la Dirección de Innovación Educativa. La figura 2 ilustra el circuito de esta primera ronda de juicio de expertos.

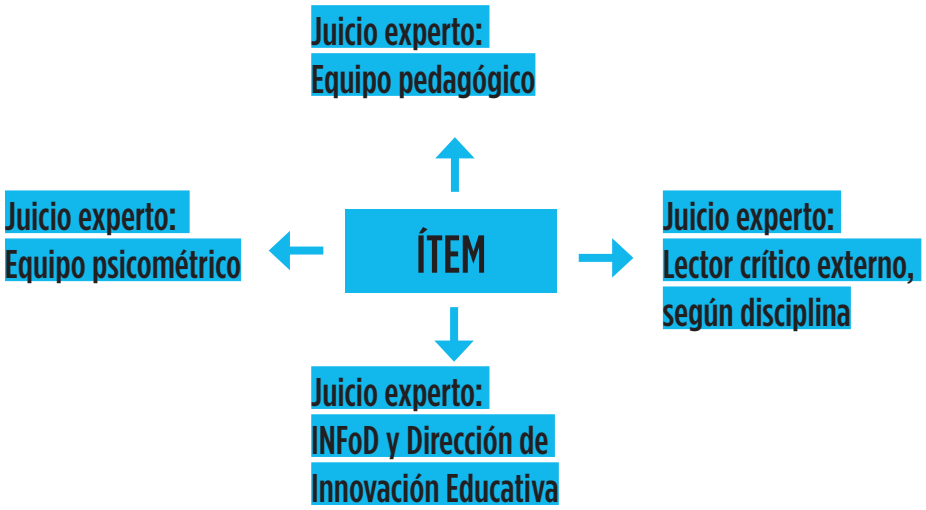

Figura 2. Juicio de expertos en la elaboración de los ítems.

Los expertos en psicometría verificaron que la estructura del ítem fuera adecuada. Se examinó que la longitud de los enunciados fuera acorde a los parámetros recomendados, que las opciones de respuesta fueran similares en su nivel de dificultad, que fueran factibles, entre otros aspectos.

Por su parte, los lectores críticos y profesionales del INFoD y de la Dirección de Innovación Educativas analizaron la claridad del enunciado y el grado de ajuste del ítem a los contenidos y capacidades enunciadas en los NAP. En cada una de estas instancias, se hicieron sugerencias para la mejora de los ítems que conformaron la versión final de las pruebas piloto.

Cuerpo colegiado de lectores críticos

Con el objetivo de fortalecer la validez de contenido de la prueba, se conformó un cuerpo colegiado de lectores críticos que evaluaron la calidad de los ítems de cada una de las áreas.

Para la conformación de este cuerpo colegiado de lectores críticos de carácter federal y regional, se hizo una convocatoria a través de la Unidades de Evaluación Jurisdiccionales (UEJ). Los expertos disciplinares debían reunir los siguientes requisitos: (a) profesionales con título de maestría o doctorado (o en curso) en las áreas a desempeñarse; (b) ejercicio de la profesión docente en el nivel primario pertenecientes a las áreas a desempeñarse; (c) capacidad de trabajar en forma autónoma en base a un cronograma; (d) disponibilidad para capacitarse y; (e) se valorará la experiencia en investigación y publicaciones en las áreas en cuestión. Se priorizó la representatividad federal regional.

Una vez seleccionados los perfiles, los especialistas de Matemática y Lengua de la SEE les enviaban una planilla de validación por ítem a leer, la cual debía ser completada y remitida a la SEE. De esta forma se fortalecieron los instrumentos de evaluación garantizando su confiabilidad y transparencia. Cada lector crítico debía leer 100 ítems de su disciplina y remitir sus comentarios teniendo en cuenta la dimensión estructural del ítem, el rigor científico de su contenido y su inclusión en los NAP y en los contenidos consensuado.

En esta evaluación participaron cuatro expertos de Matemática de las siguientes provincias: Mendoza, Misiones, Santa Fe y Buenos Aires; y cuatro expertos de Lengua de las siguientes provincias: Ciudad Autónoma de Buenos Aires, San Luis, Mendoza y Tucumán.

Este equipo trabajó sobre los ítems de primaria, realizando una lectura crítica y aportando sugerencias para el fortalecimiento del carácter federal de la prueba. Se examinó la redacción de los ítems, las opciones y claves de repuesta, los contenidos y capacidades evaluados. Básicamente, cada lector crítico debía dar su juicio considerando las siguientes características: (a) Claridad de la consigna-si se comprende fácilmente, si es adecuada sintáctica y semánticamente-; (b) Relevancia del ítem para medir la capacidad y el contenido al que se orienta; (c) Claridad de las opciones de respuestas –sintáctica y semánticamente; y (d) Adecuación de la clave y de los distractores.

Los lectores críticos indicaban sus comentarios en las Guías para la validación de ítems (ver Anexo 1) y luego, estas sugerencias eran analizadas cualitativamente por el equipo de especialistas de la SEE.

Sobre la base de este estudio de juicio de expertos se revisaron, modificaron o reelaboraron los ítems observados. De esta manera los resultados del trabajo del cuerpo colegiado de lectores críticos proveyeron de un insumo fundamental para el logro de cada vez mejores pruebas. A partir de estos datos, se conformaron las versiones finales de la prueba piloto Aprender 2018.

5) Estudio piloto

Como ya se ha mencionado, el objetivo de los estudios piloto es analizar en forma empírica el comportamiento de los ítems que han sido elaborados bajo estándares consensuados y sometidos, previamente, a juicio de expertos. Los estudios piloto permiten identificar ítems débiles o defectuosos, determinar la dificultad de cada uno, conocer la capacidad para discriminar entre desempeños altos y bajos, fijar qué ítems conformarán el test, el tiempo promedio de resolución y calcular las intercorrelaciones entre ítems para examinar la homogeneidad de la prueba.

Para garantizar que los resultados del estudio piloto sean confiables, es necesario que la capacitación de los aplicadores sea uniforme y que la muestra del estudio piloto sea representativa de la población en estudio.

Capacitación de los aplicadores del estudio piloto de las pruebas de Lengua y Matemática

Con el objetivo de garantizar las condiciones estandarizadas de aplicación de las pruebas piloto de Aprender 2018, la SEE –en colaboración con las jurisdicciones implicadas–, llevó a cabo diferentes jornadas de capacitación para los aplicadores. En dichas jornadas se presentaron los lineamientos generales del operativo, se familiarizó a los aplicadores con los materiales y con el formato de la prueba y se brindaron las pautas generales a mantener durante la aplicación (e.g .instrucciones y tiempo para completar la prueba). Dichas pautas fueron a su vez entregadas por escrito a través de una guía paso a paso y un documento elaborado a tal fin.

Las capacitaciones de aplicadores Aprender 2018 incluyeron información acerca de los objetivos, importancia del resguardo y confidencialidad de la evaluación, características generales de las pruebas Aprender 2018, lineamientos para las actividades con los directores de las escuelas (presentación, selección de la sección, acuerdo para tomar decisiones respecto de cuestiones prácticas de la evaluación –por ejemplo, cómo proceder con los alumnos que terminan antes las pruebas; con los alumnos en integración y con los docentes o referentes jurisdiccionales para permanecer en el aula durante la evaluación), lineamientos para las actividades con los estudiantes (presentación, orden de las evaluaciones, tiempo, uso de la calculadora, apagado y guardado de celulares y dispositivos electrónicos, materiales) y relación con los coordinadores de la SEE.

De esta manera se garantizó que los pilotos de las pruebas Aprender 2018 fueran realizados bajo las mismas condiciones por todos los estudiantes que participaron en la evaluación. Por otra parte, todos los alumnos participantes tenían los mismos instrumentos para responder los cuestionarios (ej. lápiz, goma), la misma cantidad de tiempo, y el mismo tiempo de intervalo entre las diferentes áreas evaluadas.

Estudio piloto de las pruebas Aprender 2018 de Matemática y Lengua

Con el objetivo de analizar el comportamiento psicométrico de los ítems elaborados para las pruebas Aprender 2018 se llevaron a cabo un estudio piloto durante el mes junio del 2018. En cada uno, se pusieron a pruebas distintos conjuntos de ítems de las cuatro áreas evaluadas en los diferentes niveles.

En todos los casos las pruebas fueron realizadas en papel y completadas con lápiz. Tal como se ha señalado, la mayoría de las pruebas Aprender incluyen ítems de opción múltiple con 4 opciones de respuesta. En las pruebas piloto de Aprender, se utilizaron cuadernillos que tenían 30 ítems por área. Todos los alumnos contaron con aproximadamente 60 minutos más 10 minutos extra (70 minutos en total) para contestar cada una de las áreas y 10 minutos de receso entre una y otra área evaluada.

El piloto se realizó durante los días 27 y 28 de junio de 2018. Se evaluaron 1546 estudiantes pertenecientes a 59 secciones de 59 escuelas distribuidas en cinco jurisdicciones: CABA, Provincia de Buenos Aires (Mar del Plata y Gran Buenos Aires), Córdoba (Capital), Salta (Capital, General Mosconi y Tartagal), Mendoza (Capital y Guaymallén). Las áreas y años incluidos en la prueba fueron: Matemática y Lengua en 6° año de la Educación Primaria.

Tabla 5. *Participación prueba piloto Aprender 2018*[2]

Total	
Escuelas participantes	59
Secciones participantes	59
Alumnos participantes	1546
Lengua	
Escuelas participantes	58
Secciones participantes	58
Alumnos participantes	1201
Matemática	
Escuelas participantes	59
Secciones participantes	59
Alumnos participantes	1223

Cada estudiante de la Educación Primaria contestó dos evaluaciones: Matemática y Lengua y un cuestionario complementario con preguntas referidas al contexto sociodemográfico. Las pruebas de Matemática y Lengua incluyeron cada 30 ítems de opción de respuesta múltiple y los estudiantes disponían de 60 minutos para contestar cada una de las pruebas más 10 minutos extra (70 minutos en total).

En el área de Matemática en 6° año de la Educación Primaria se evaluaron 8 modelos conformados por 2 Bloques de 15 ítems cada uno (8 bloques - 8 modelos).

En el área de Lengua en 6° año de la Educación Primaria se evaluaron 8 modelos conformados por 2 Bloques de 15 ítems cada uno (8 bloques - 8 modelos).

Análisis de los resultados de las pruebas piloto

El enfoque general para el análisis de los resultados toma en cuenta la TRI, específicamente el modelo de dos parámetros que permite, para cada ítem, estimar no sólo su dificultad sino también su capacidad de discriminación (SEE, 2017). Este modelo de dos parámetros requiere de un mayor número de casos que el de un parámetro. Por ello y dado que se cuenta con una muestra para la prueba piloto Aprender 2018 (mientras que en la prueba definitiva Aprender 2018 se trata de un estudio censal), para el procesamiento de los datos del piloto, se utilizó un modelo de un solo parámetro, mientras que para el procesamiento de los datos de la prueba definitiva se utilizó un modelo de dos parámetros.

2 Se considera escuela participante a aquellas que tengan al menos un alumno que haya hecho alguna marca en algún cuestionario.
Se considera sección participante a aquellas que tengan al menos un alumno que haya hecho alguna marca en algún cuestionario.
Se considera alumno participante a aquellos que tengan al menos una marca en algún cuestionario.

Los datos del estudio piloto fueron procesados por el equipo metodológico quienes se encargaron del diseño y análisis de datos. Se calcularon los índices de dificultad relativa de los reactivos y, con el objeto de analizar la homogeneidad de los ítems en su conjunto, el coeficiente de correlación biserial.

Para el análisis de los resultados de las dos áreas (Matemática y Lengua 6° año de la Educación Primaria) se consideró la distribución del nivel de dificultad de los ítems (que todo el continuo de dificultad estuviera adecuadamente representado), la correlación biserial y la distribución de las respuestas en las cuatro opciones correspondientes del cada ítem.

En cuanto al coeficiente de correlación biserial, se consideró como indicador psicométrico satisfactorio un valor ≥ 0.30, aunque en algunos ítems se registraron coeficientes más bajos. No obstante, para eliminar un ítem se consideró no solo la correlación biserial sino también otros indicadores psicométricos (e.g. el índice de dificultad del ítem). En función de ello, se seleccionaron los ítems con mejor comportamiento psicométrico y se eliminaron los ítems anómalos, vale decir aquellos que mostraran valores inadecuados en los indicadores paramétricos- cuidando que el contenido específico del ítem eliminado quedara representado por otros ítems en la totalidad de la prueba y que la dificultad del ítem eliminado no generara una "laguna" en el rango de dificultad de la prueba.

En el Anexo 2 se presentan los histogramas con la distribución de los ítems según nivel de dificultad para cada una de las áreas y años evaluados en la prueba definitiva Aprender 2018. También se muestra la correlación entre la dificultad de los ítems en el piloto y la obtenida en la prueba definitiva

Efecto del orden en que se toman las evaluaciones en la participación

Uno de los objetivos de la prueba piloto de Aprender 2018 fue analizar el efecto del orden de presentación de los materiales sobre la tasa de respuesta. En consonancia, las pruebas fueron administradas en forma contrabalanceada. Es decir, se conformaron dos grupos de estudiantes, un grupo contestó primero la prueba de Lengua y luego la prueba de Matemática; y otro, en forma inversa.

Para evaluar el efecto del orden de presentación se utilizaron dos indicadores: (a) la tasa de alumnos que respondieron al menos un ítem de la prueba y la abandonaron en medio de su realización, dejando sin contestar, al menos, las últimas cinco preguntas y; (b) la media de ítems sin responder por alumno en cada materia. Al primer indicador se lo denominó "abandono" y al segundo "no respuesta".

En la figura 3 se presentan los resultados obtenidos en cada área para el indicador "abandono" (tasa de alumnos que dejaron la prueba incompleta).

Efecto del orden de las evaluaciones en el abandono de la prueba, según año y materia

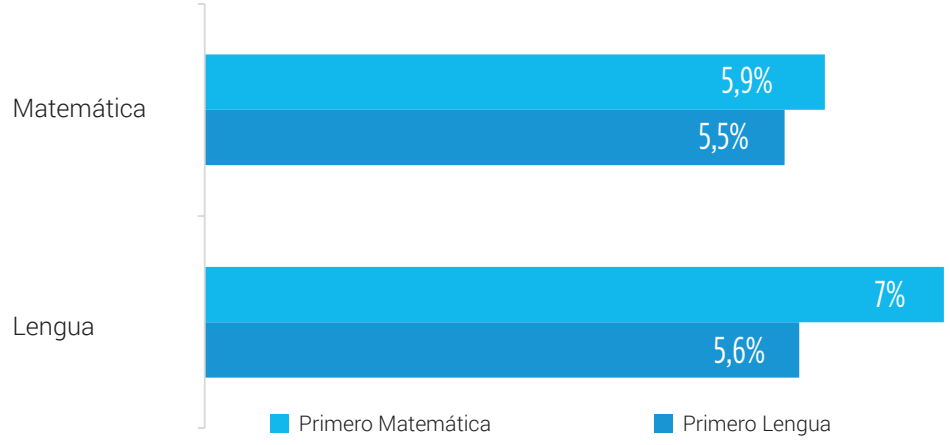

Figura 3. Efecto del orden de las evaluaciones en el "abandono" de la prueba según materia.

En la figura 4, se presentan los resultados obtenidos para cada área para el indicador "no respuesta" (promedio de ítems no respondidos por estudiante).

Efecto del orden de las evaluaciones en la no respuesta, según año y materia

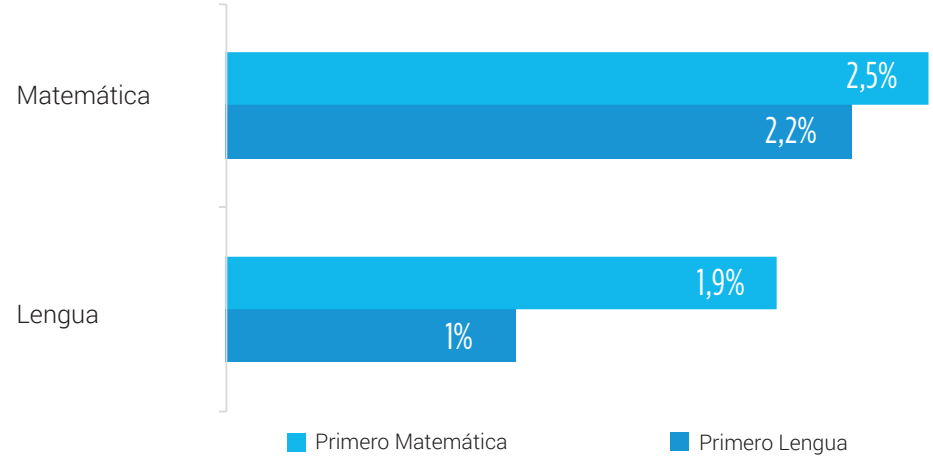

Figura 4. Efecto del orden de las evaluaciones en la "no respuesta" según materia.

Tal como se puede observar para ambos indicadores, los valores son mayores cuando la prueba resuelta en primer lugar fue la de matemáticas.

Para evaluar si estas diferencias eran significativas a nivel estadístico se utilizó la Prueba t de Student para muestras independientes. No se encontraron diferencias estadísticamente significativas en ninguno de los dos indicadores. Sin embargo, tal como se puede observar en las figuras 3 y 4, hay una direccionalidad clara en los resultados encontrados. En efecto, tanto el abandono como la no respuesta, aumentan al evaluar Matemática en primer lugar. De acuerdo con ello, si se administra primero Lengua la probabilidad de que aumenten las respuestas es superior a que si se aplica primero Matemática. Además, dado que en operativos anteriores se utilizó el orden Lengua-Matemática, se consideró que este modo de presentación suponía una mayor comparabilidad con los estudios precedentes y se recomendó seguir este orden en la prueba definitiva.

6) Revisión de las pruebas y preparación de las formas definitivas de las pruebas Aprender 2018

A partir de los resultados de las pruebas piloto se seleccionaron los ítems que conformarían las pruebas Aprender 2018 para cada una de las áreas y año evaluados. En la selección de los ítems se tuvieron en cuenta los siguientes criterios:

- Aprobación de especialistas del equipo pedagógico de la SEE
- Aprobación de lectores críticos internos
- Aprobación por juicio de expertos externo y federal
- Aprobación por especialistas en INFoD
- Aprobación por especialistas en psicometría
- Indicadores psicométricos satisfactorios

A continuación, se presentan las tablas de especificaciones en donde se detallan la cantidad y porcentaje de ítems según capacidad y contenidos para cada una de las áreas y año evaluados.

Tabla 6. *Tabla de especificación de Matemática 6° año de la Educación Primaria*

Capacidad Contenido	Reconocimiento de conceptos	Solución de operaciones	Resolución de problemas	Comunicación en matemática	Total
Número y operaciones	5 ítems 7%	5 ítems 7%	19 ítems 26%	9 ítems 13%	38 ítems 53%
Geometría y medida	8 ítems 11%		13 ítems 18%		21 ítems 29%
Estadística y probabilidad			8 ítems 11%	5 ítems 7%	13 ítems 18%
Total	13 ítems 18%	5 ítems 7%	40 ítems 56%	14 ítems 19%	72 ítems 100%

Tabla 7. *Tabla de especificación de Lengua 6° año de la Educación Primaria*

Contenidos	Extraer información explícita	Interpretar información sugerida	Reflexionar y evaluar sobre distintos aspectos textuales	Total
Aspectos globales del texto (tema, secuencia, narrador, etc.	2 ítems 2.8%	13 ítems 18.1%	14 ítems 19.4%	29 ítems 40.3%
Aspectos locales del texto (cohesión, vocabulario, información explícita, etc.)	19 ítems 26.4%	22 ítems 30.5%	2 ítems 2.8%	43 ítems 59.7%
% Total	21 ítems 29.2%	35 ítems 48.6%	16 ítems 22.2%	72 ítems 100%

Por último, para analizar la confiabilidad de las pruebas definitivas Aprender 2018 se utilizó el método de consistencia interna basado en el coeficiente de alfa de Cronbach. En la tabla 8 se muestran los coeficientes de alfa de Cronbach para cada uno de los modelos de Matemática y de Lengua. Tal como se puede observar, en el caso de Matemática los valores de coeficiente de alfa de Cronbach variaron entre .79 y .81 y, en el caso de Lengua, los valores fueron de .77 a .82. Para la interpretación de los resultados se tomó como criterio general las recomendaciones propuestas por George y Mallery (2003). Ellos sugieren que valores de alfa mayores a .90 son considerados excelentes; entre .80 y .90, muy buenos; mayores que .70 son buenos; entre .70 y .60 cuestionables; entre .60 y .50 son débiles; y los menores a .50 resultan inaceptables. En este sentido, los resultados obtenidos aportan evidencias satisfactorias de consistencia interna.

Tabla 8. *Resumen de resultados del análisis de consistencia interna pruebas Aprender 2018*

Modelo	Alfa de Cronbach	
	Matemática	Lengua
1	.82	.81
2	.79	.79
3	.80	.81
4	.80	.82
5	.81	.77
6	.82	.82

CONCLUSIONES

Una evaluación es una tarea que sirve para conocer un ámbito, una situación, un área o un dominio y tiene por objetivo orientar la toma de decisiones a partir de los conocimientos obtenidos y beneficiar a los actores involucrados en dicho ámbito.

La evaluación educativa tiene el propósito principal de otorgar datos que contribuyan al fortalecimiento de las prácticas de enseñanza y a la planificación de estrategias de mejora de los aprendizajes. Para cumplir con este objetivo es menester contar con instrumentos válidos y confiables y documentación que respalde y dé cuenta de los procesos que han sido llevados a cabo para obtener dichos instrumentos y verificar su funcionamiento.

Cada proceso de evaluación educativa involucra distintas etapas. Una de las primeras y fundamentales es el diseño y construcción de las pruebas con las que se evaluará a los estudiantes. Como todo instrumento de medida, las pruebas tienen que pasar por controles de calidad en cuanto a su construcción para garantizar su uso idóneo. Las evidencias en cuanto a la validez y fiabilidad son cruciales ya que de ello depende que las interpretaciones y conclusiones llevadas a cabo en un análisis posterior sean representativas del dominio evaluado y sustantivos para la descripción del desempeño de los estudiantes.

Este documento tiene como propósito esencial comunicar cómo han sido construidas las pruebas Aprender 2018, brindar algunas evidencias acerca de su validez y confiabilidad y dar a conocer todos los recaudos y consideraciones metodológicas que se han tenido en cuenta en el proceso de construcción y diseño de la prueba a fin de garantizar resultados significativos y psicométricamente adecuados.

Entendemos que la mejora de los procesos de aprendizaje y de enseñanza depende de la articulación y del trabajo conjunto y colaborativo de los diferentes actores y sectores del sistema. Esperamos que la información aquí brindada sirva al entendimiento de nuestra tarea y nuestro compromiso con el sistema educativo argentino.

REFERENCIAS

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (2014). *Standards for educational and psychological testing*. Washinton, DC: American Psychological Association

Cortada de Kohan, N., Macbeth, G., & López Alonso, A. (2008). *Técnicas de Investigación Científica*. Buenos Aires, Argentina: Lugar.

Covacevich, C. (2014). *Cómo seleccionar un instrumento para evaluar aprendizajes estudiantiles. Nota Técnica*. Banco Interamericano de Desarrollo. División de Educación (SCL/EDU). Obtenido el 29 de noviembre de 2017: https://publications.iadb.org/bitstream/handle/11319/6758/C%C3%B3mo-seleccionar-un-instrumento-para-evaluar-aprendizajes-estudiantiles.pdf

Cronbach, L. J. (1971): *Fundamentos de la exploración psicológica*. Madrid: Biblioteca Nueva.

George, D. & Mallery, P. (2003). *SPSS/PC+ step by step: A simple guide and reference*. Belmont, CA: Wadsworth Publishing Company.

Glaser, R. (1963). *Instructional technology and the measurement of learning out-comes: Some questions*. American Psychologist, 18, 519-521.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory (Vol. 2)*. California: Sage.

Hambleton, R.K. & Swaminathan, H. (1978). *Criterion-Referenced Testing and Measurement: A review of technical issues and developments*. Review of Educational Research, 40, 1-47.

Jornet Meliá, J. M. & González Such, J. (2009). *Evaluación criterial: determinación de estándares de interpretación (EE) para pruebas de rendimiento educativo*. Estudios sobre Educación, 16, 103-123.

Leones, M. (2005). *Recorrido político y técnico-pedagógico en el proceso de elaboración, justificación y validación de los criterios de evaluación*. Ministerio de Educación, Ciencia y Tecnología de la Nación. Dirección Nacional de Información y Evaluación de la Calidad Educativa (DINIECE). Obtenido el 29 de noviembre de 2017: http://repositorio.educacion.gov.ar:8080/dspace/bitstream/handle/123456789/95653/evaluacion_criterial_6abril06.pdf?sequence=1

Leyva Barajas, Y. E. (2011). *Una reseña sobre la validez de constructo de pruebas referidas a criterio*. Perfiles Educativos, 33(131), 131-154.

Livingston, S. A. (2009). *Constructed-Response Test Questions: Why we use them; How we score them*. Obtenido el 29 de Noviembre de 2017 en: https://www.ets.org/Media/Research/pdf/RD_Connections11.pdf

Ministerio de Educación, Ciencia y Tecnología. (2001). *Recorrido político y técnico –pedagógico en el proceso de elaboración, justificación y validación de los criterios de evaluación*. Dirección Nacional de Información y Evaluación de la Calidad Educativa – DiNIECE. Obtenido el 14 de Febrero de 2019 en: http://www.bnm.me.gov.ar/giga1/documentos/EL001414.pdf

Martínez Arias, M. R. (1995). *Psicometría: Teoría de los test psicológicos y educativos*. Madrid: Síntesis.

Muñiz, J. (2010). *Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems*. Papeles del Psicológico, 3(1), 57-66.

Secretaría de Evaluación Educativa (2017). Apender 2016. *Notas Metodológicas. Serie de Documentos Técnicos/2*. Ministerio de Educación y Deportes de la Nación. Obtenido el 30 de Noviembre de 2017 en: http://www.bnm.me.gov.ar/giga1/documentos/EL005591.pdf

ANEXO

ANEXO I

Guía para la validación de ítems – Matemática 6° Año Educación Primaria

Lector crítico: .....

Código del Ítem :			
Autor:	**Si**	**No**	**Sugerencias**
1.¿La clasificación del contenido es correcta?			
2.¿La capacidad cognitiva es correcta?			
3.¿Responde a uno de los contenidos consensuados?			
4. ¿Es apropiado al grado o año?			
5.¿Puede tener algún sesgo cultural o de género o molestar a un grupo?			
6.¿El cuerpo es claro y preciso en su redacción?			
7.¿La clave es correcta?			
8.¿Los distractores son plausibles?			
9.¿Se puede llegar a la clave por un método equivocado?			
10.¿Responden las alternativas a un mismo campo conceptual y son similares en apariencia?			
11.¿Las alternativas están ordenadas y encolumnadas por las unidades?			
12.¿Los gráficos tienen título y nombre en los ejes?			
13.¿Se usan expresiones negativas?			
14.¿Se repiten palabras en los distractores?			
15.¿Está redactado en tercera persona?			
16.¿Las alternativas son sencillas y cortas?			
17.¿Están correctamente justificados los distractores?			
18.¿Están redactadas correctamente las hipótesis de error?			

Guía para la validación de ítems – Lengua 6° Año Educación Primaria

Se espera que el lector crítico provea información de relevancia en relación con los siguientes aspectos de los ítems de respuesta múltiple

Dimensión teórica

Contempla el rigor disciplinar de los contenidos que son objeto de evaluación del ítem, de manera de corroborar que el mismo no contiene ni encubre errores.

Dimensión epistemológico -didáctica

Contempla:

- Relevancia. Está determinada por el nivel de representatividad del objeto de evaluación en los Núcleos de Aprendizaje Prioritarios (NAP) y por los marcos teóricos allí explicitados. La relevancia responde a las preguntas: ¿es este contenido pertinente de ser evaluado en las escuelas de toda la República Argentina?, ¿está incluido en los NAP? (recordar que se evalúa terminalidad; es decir contenidos trabajados en los 3 últimos años de Educación Primaria).

- Transposición didáctica. Está determinada por la significatividad lógica y psicológica al nivel educativo en el que será utilizado el ítem. Responde a la pregunta ¿puede ser contestado por estudiantes de X nivel educativo de las escuelas de todo el país?

- Adecuación lingüística. Está determinada tanto por el estilo discursivo y su relación con el nivel educativo en el que se evalúa como por la adecuación al léxico propio de la disciplina. Responde a las preguntas ¿contiene términos polisémicos?, ¿el lenguaje puede ser entendido por estudiantes de X nivel educativo de las escuelas de todo el país?, ¿contiene términos específicos innecesarios para la resolución del ítem?, entre otras.

Dimensión estructural

Contempla:

- Adecuación del enunciado. Está determinada por la presentación de un contexto relevante, posible y necesario para la resolución de la consigna. Responde a las preguntas: ¿es verosímil y coherente?, ¿presenta toda la información necesaria para la resolución del ítem?, entre otras.

- Precisión de la consigna. Está dada por la estructura lingüística que determina la claridad de la misma, de manera que los estudiantes comprendan sin lugar a dudas la tarea a resolver. Responde a preguntas como ¿está formulada en positivo?, ¿incluye solo un verbo de acción?, ¿puede ser contestada sin utilizar la información incluida en las opciones de respuesta?

- Precisión de la respuesta correcta. La clave o respuesta correcta debe ser la única de las 4 opciones que responde de manera correcta y completa a la consigna.

- Adecuación de las opciones de respuesta incorrectas. Está determinada por la información presentada en los 3 distractores. Responde a las preguntas: ¿son todas las opciones incorrectas pero plausibles y verosímiles?, ¿son todas posibles respuestas a la consigna?, ¿alguna de las opciones es parcialmente correcta? (esto no puede ocurrir), ¿responden todas a la misma dimensión conceptual?

Para realizar la validación deberán llenar por cada ítem el cuadro que se encuentra a continuación. En la columna Validación deberán escribir validado o explayarse en los motivos por los cuales no se valida un determinado aspecto del ítem y proponer los cambios a realizar, cuando estos sean mínimos. Por ejemplo, cambiar términos, modificar la redacción, etc.

Código de materia (L6) (__n° de ítem):

Elaborador del ítem:

Texto de referencia: (Título)

Lector externo:

Criterios	Sí	No	Sugerencias
Calidad del Ítem			
1-¿Es clara la tarea asignada a los alumnos?			
2- ¿El contenido conceptual es académicamente correcto?			
3-¿El contenido está identificado correctamente?			
4-¿La capacidad cognitiva está señalada correctamente?			
5-¿Hay una sola respuesta correcta?			
6-¿El lenguaje utilizado es apropiado a la edad?			
7-¿El ítem contiene prejuicios culturales, de género, etc.?			
Distractores			
8-¿Los distractores son creíbles aunque se demuestren claramente incorrectos?			
9-¿Los distractores están en orden (cronológico, de menor a mayor o de mayor a menor)?			
10-¿Las estructuras gramaticales y la dimensión léxica de los distractores son equivalentes?			
11-¿Se incluyeron en los distractores ideas			
alternativas o previas, o conceptos erróneos y errores comunes o confusiones habituales de los alumnos?			

Normativa y gramática			
12-¿Se ha respetado la normativa sobre la puntuación en el ítem directo?			
13-¿Se ha respetado la normativa sobre la puntuación en el ítem de completamiento?			
14-¿Hay un uso preciso de verbos, adverbios y adjetivos?			
15- ¿Las palabras o expresiones destacadas están subrayadas o resaltadas en negrita?			
16-¿Las citas textuales están en cursiva?			

Fecha de devolución:

Serán rechazados en esta instancia los ítems que tengan una calificación negativa en los apartados 1, 2, 3 y 4.

Se invalidará la totalidad de ítems de aquellos elaboradores que en esta primera etapa presenten un 40% de ítems con una calificación negativa en los apartados 1, 2, 3 y 4.

ANEXO 2

En las figuras 5 y 6 se muestra la distribución de los ítems según nivel de dificultad para cada una de las áreas y años evaluados en Aprender 2018. Se observa que el rango de ítems de dificultad baja se encuentra representado (entre -2 y -1), aunque en todos los histogramas hay una leve asimetría hacia la derecha, lo que indica una mayor cantidad de ítems de dificultad mayor.

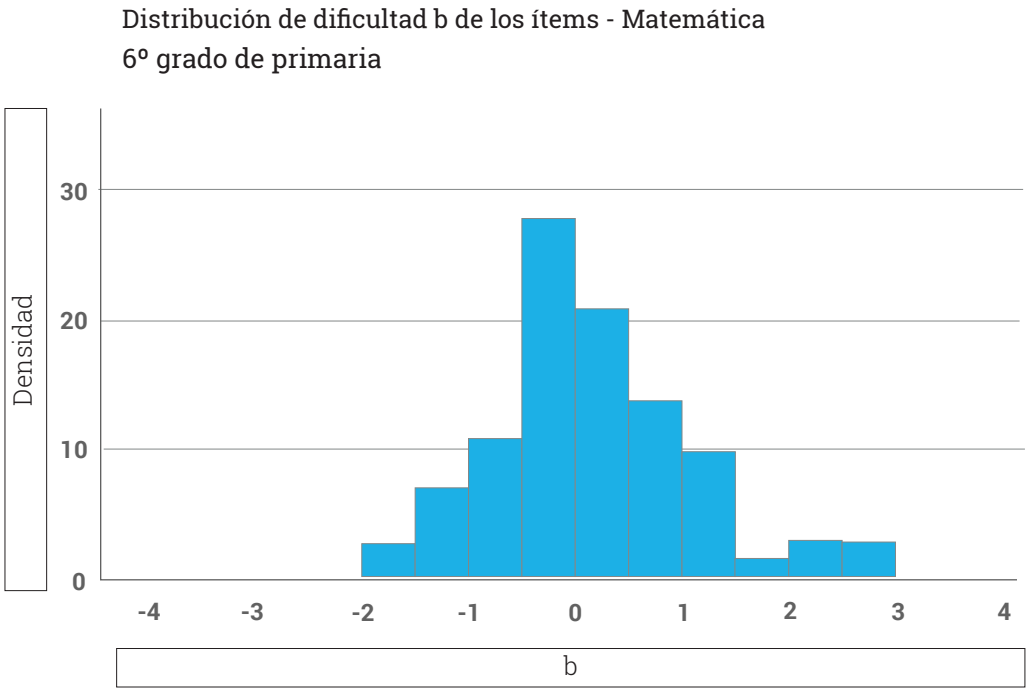

Figura 3. Histograma de la dificultad: Matemática 6° Año de la Educación Primaria

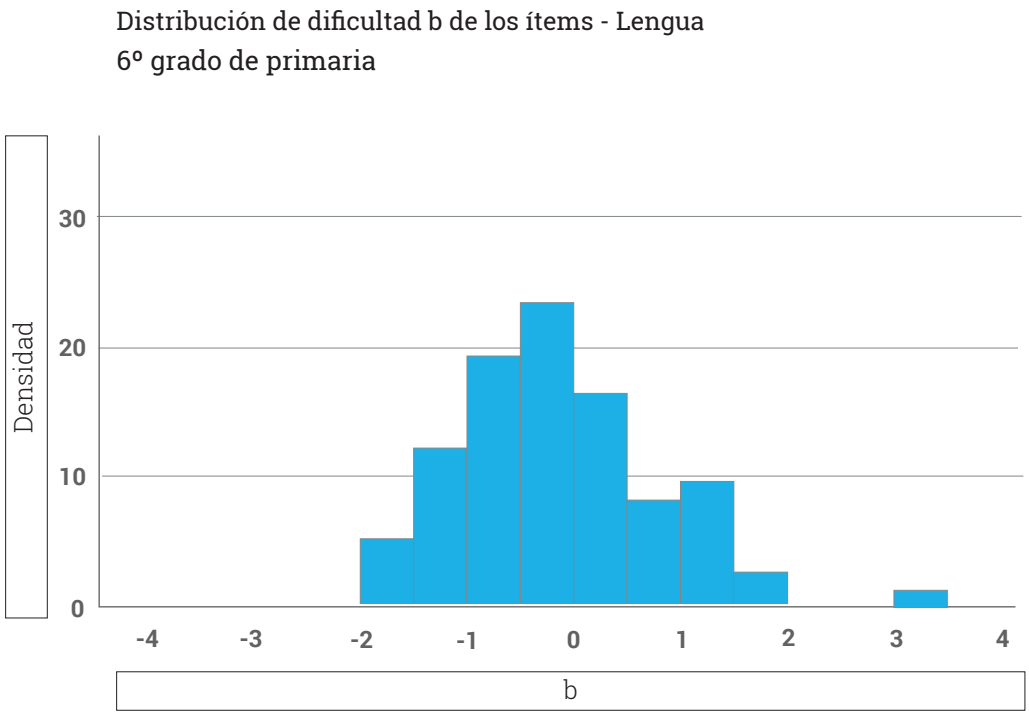

Figura 4. Histograma de la dificultad: Lengua 6° Año de la Educación Primaria

En las figuras 3, 4, 5 y 6 se muestra la distribución de los ítems según nivel de dificultad para cada una de las áreas y años evaluados en Aprender 2017. Se observa que el rango de ítems de dificultad baja se encuentra representado (entre -2 y -1), aunque en todos los histogramas hay una leve asimetría hacia la derecha, lo que indica una mayor cantidad de ítems de dificultad mayor.

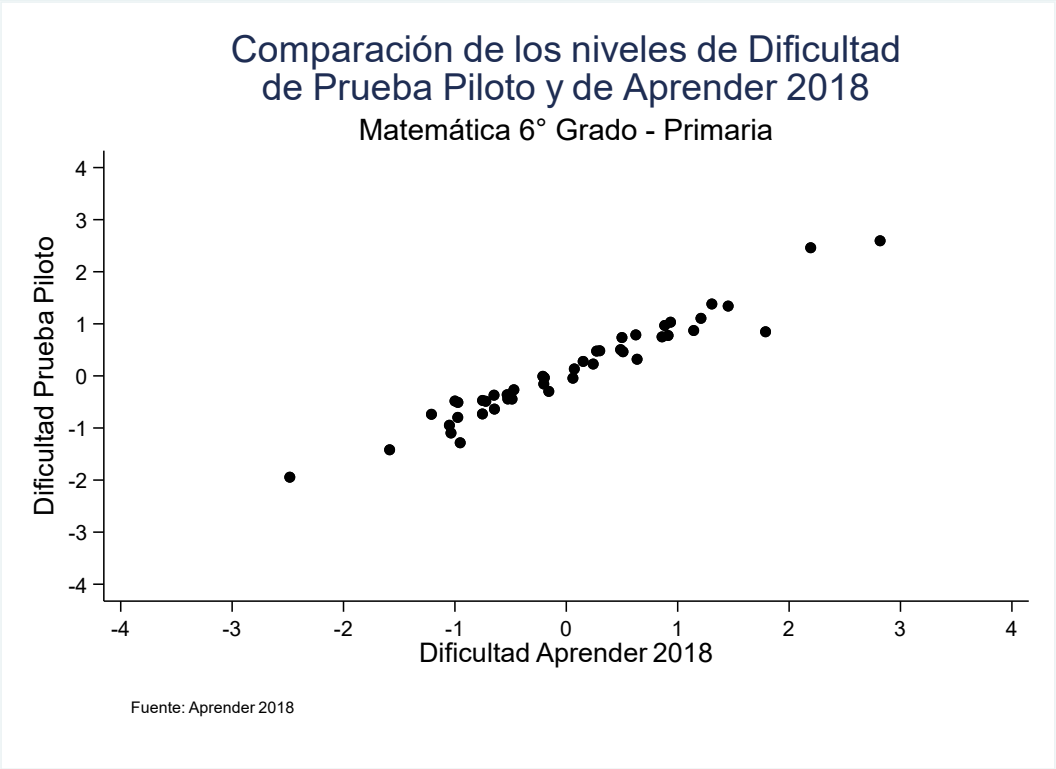

Figura 5. Comparación de la distribución de la dificultad del piloto y de la prueba definitiva Aprender 2018 de Matemática 6° año – Nivel Primario.

En los diagramas de dispersión que se muestran a continuación, se puede observar la correlación entre la dificultad de los ítems en las pruebas piloto y la dificultad obtenida en la prueba definitiva. Estos gráficos evidencian una alta correlación en la dificultad observada por ambas evaluaciones en todas las áreas y años incluidos.

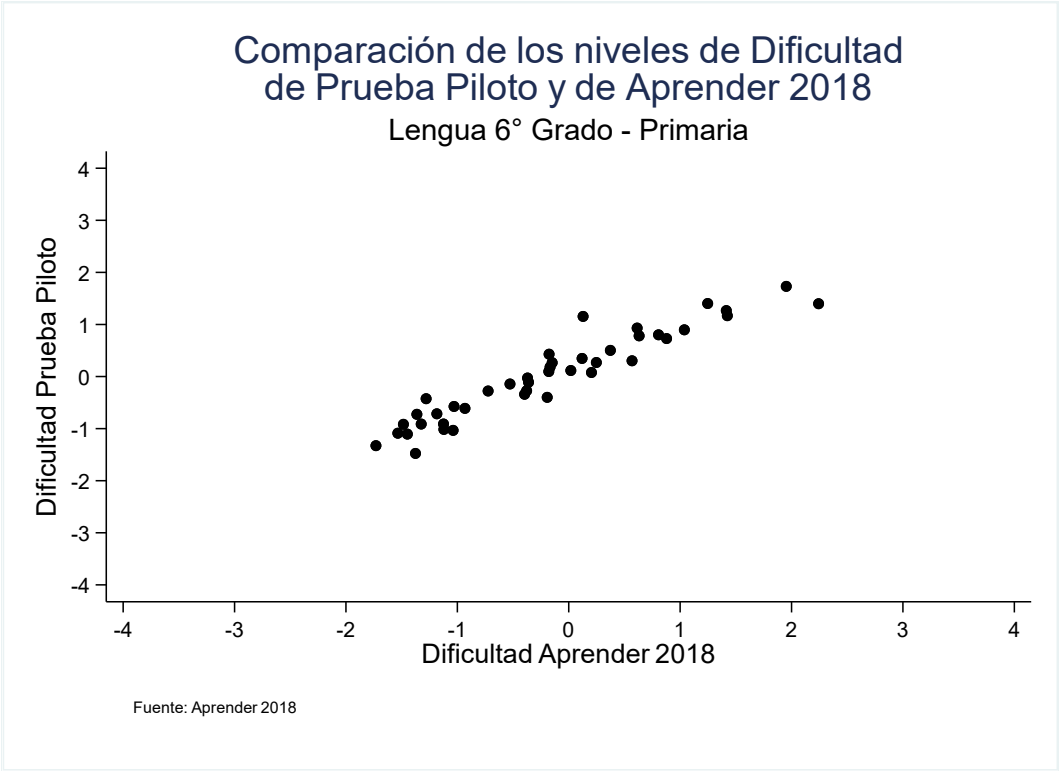

Figura 6. Comparación de la distribución de la dificultad del piloto y de la prueba definitiva Aprender 2018 de Lengua 6° año – Nivel Primario.

Marzo 2019,
Ciudad Autónoma de Buenos Aires

Aprender

Secretaría de
Evaluación Educativa

Ministerio de Educación,
Cultura, Ciencia y Tecnología
Presidencia de la Nación