SERIE
REPORTES
TÉCNICOS

COMPETENCIA
LOGROS
LECTURA
MATEMÁTICA
MEDICIÓN
MUESTREO
DESEMPEÑO
PSICOMETRÍA
APRENDIZAJE
EDUCACIÓN
CALIDAD

Reporte técnico de la Evaluación Muestral de Estudiantes (EM) 2018 2.° grado de primaria

PERÚ Ministerio de Educación

El Perú Primero

SERIE
REPORTES
TÉCNICOS

COMPETENCIA
LOGROS
LECTURA
MATEMÁTICA
MEDICIÓN
MUESTREO
DESEMPEÑO
PSICOMETRÍA
APRENDIZAJE
EDUCACIÓN
CALIDAD

Reporte técnico de la Evaluación Muestral de Estudiantes (EM) 2018 2.° grado de primaria

PERÚ Ministerio de Educación

El Perú Primero

PERÚ | Ministerio de Educación

Flor Aidee Pablo Medina
Ministra de Educación del Perú

Guido Alfredo Rospigliosi Galindo
Viceministro de Gestión Institucional

Ana Patricia Andrade Pacora
Viceministra de Gestión Pedagógica

José Carlos Chávez Cuentas
Secretario de Planificación Estratégica

Humberto Pérez León Ibáñez
Jefe de la Oficina de Medición de la Calidad de los Aprendizajes

Reporte técnico de la Evaluación Muestral de Estudiantes (EM) 2018
2.° grado de primaria

Responsables del documento
Andrés Burga León
Gabriela Santibañez Rojas
Giovanna Moreano Villena
Luis Mejía Campos
Tania Pacheco Valenzuela
Yuriko Sosa Paredes

Esta publicación es el producto final del esfuerzo institucional de la UMC por medio de sus diferentes equipos de especialistas.

Calle Del Comercio 193, San Borja
Lima, Perú
Teléfono: (511) 615-5800
www.minedu.gob.pe

En el presente documento, se utilizan de manera inclusiva términos como “el docente”, “el estudiante” y sus respectivos plurales (así como otras palabras equivalentes en el contexto educativo) para referirse a hombres y mujeres. Esta opción se basa en una convención idiomática y tiene por objetivo evitar las formas para aludir a ambos géneros en el idioma castellano (“o/a”, “los/las” y otras similares), debido a que implican una saturación gráfica que puede dificultar la comprensión lectora.

Índice

Introducción

La Oficina de Medición de la Calidad de los Aprendizajes (UMC) de la Secretaría de Planificación Estratégica del Ministerio de Educación (Minedu) tiene entre sus funciones el diseño e implementación de las evaluaciones de logros de aprendizaje de los estudiantes de Educación Básica a nivel nacional. En ese marco, y de acuerdo a la RM 116-2018-Minedu, el año 2018 la UMC llevó a cabo la Evaluación Muestral de estudiantes (EM) en 2.° grado de primaria. A partir de ella, se recogió información sobre las áreas de Matemática (Resuelve problemas de cantidad) y Comunicación (Lectura), la cual tiene representatividad nacional y por estratos (gestión, área y característica). El presente documento brinda información técnica de esta evaluación.

En el primer capítulo, se describen la población objetivo y el marco muestral utilizado. Se señala cómo fue definido el tamaño de la muestra y los pesos utilizados en la estimación de los diversos parámetros derivados de dicha muestra. Además, se añade información sobre la validación de la muestra de la EM 2018.

El segundo capítulo aborda los aspectos relacionados con el operativo de campo. Describe los controles de calidad, el modo cómo fueron seleccionados y capacitados los aplicadores, y la manera cómo se recolectaron los datos en el contexto de la EM 2018.

Finalmente, el tercer capítulo describe el modelo Rasch utilizado para los análisis psicométricos. Se pone énfasis en la forma cómo se analizaron los ítems, y cómo se obtuvieron evidencias de confiabilidad y validez, análisis de precisión y consistencia de la asignación a los niveles de logro. Además, se describe cómo se realizó el proceso de equiparación de medidas.

Población y muestra

Capítulo 1

Capítulo 1

Población y muestra

1.1. Población objetivo

El marco poblacional de la EM 2018 de 2.° grado de primaria mantuvo el mismo criterio empleado para la ECE desde el año 2007: incluir solo a las escuelas que tienen cinco o más estudiantes en el grado a evaluar[1]. Sin embargo, a diferencia de los años anteriores, la evaluación de este grado en 2018 se realizó a través de una muestra de estudiantes a nivel nacional. Como en todo proceso de muestreo, se siguieron procedimientos que buscan asegurar que los resultados de la EM 2018 sean representativos de la población de 2.° grado de primaria.

En este caso, el muestreo empleado fue de tipo probabilístico, estratificado y por conglomerados de dos etapas donde la unidad primaria de muestreo corresponde a las instituciones educativas y la unidad secundaria corresponde a las secciones completas dentro de cada IE. En la primera etapa, la selección de escuelas fue proporcional al tamaño, por lo cual las escuelas con mayor cantidad de estudiantes tuvieron una mayor probabilidad de ser seleccionadas. En la segunda etapa, en las IE que tienen tres o más secciones, se eligieron dos de ellas al azar[2]. Esta selección de las secciones fue aleatoria simple, por lo que cada una tuvo la misma probabilidad de ser elegida para conformar la muestra[3].

Además, el muestreo fue estratificado porque, en un contexto de tanta diversidad como el peruano, la variable de interés (rendimiento) toma diferentes valores en diferentes subpoblaciones (área, gestión o característica), por lo que es importante reportar resultados para cada una de ellas. La combinación de estas tres subpoblaciones permite generar los siguientes estratos mutuamente excluyentes y colectivamente exhaustivos donde cada elemento pertenece únicamente a un estrato.

Estrato 1: Urbano – Estatal – Polidocente Completo.
Estrato 2: Urbano – Estatal – Unidocente / Multigrado.
Estrato 3: Rural – Estatal – Polidocente Completo.
Estrato 4: Rural – Estatal – Unidocente / Multigrado.
Estrato 5: No estatal (Las IE no estatales, casi en su totalidad, son urbanas y polidocentes).

[1] El porcentaje de estudiantes excluidos por este criterio oscila alrededor del 5 % en primaria.
[2] El porcentaje de escuelas de primaria con más de dos secciones en la muestra fue de 29,6 %.
[3] El factor de expansión, el cual permite expandir la muestra a la población, considera las dos etapas de selección señaladas.

Una vez estratificado el marco muestral, se empleó la asignación de Neyman para determinar el tamaño de cada uno de los estratos, los cuales son directamente proporcionales a la variabilidad y tamaño del mismo. La tabla 1.1 presenta el tamaño final de la muestra y de los estratos que la componen.

Tabla 1.1. ***Cantidades de IE y estudiantes por estrato en la EM 2018***

Estrato	IE	Estudiantes
Estatal Urbano Polidocente Completo	133	6 507
Estatal Urbano Unidocente / Multigrado	14	205
Estatal Rural Polidocente Completo	19	366
Estatal Rural Unidocente / Multigrado	93	873
No estatal	112	3 351
Total	371	11 302

Respecto a los niveles de inferencia de la EM 2018 en 2.° grado de primaria, se pueden reportar los siguientes resultados representativos: a) nivel nacional, b) por sexo de los estudiantes, c) por gestión de las IE (estatales / no estatales), d) por ubicación geográfica de las IE (urbanas / rurales), e) por característica (polidocente completo / unidocente-multigrado). Cabe notar que, a diferencia de las muestras de control de las ECE[4], esta muestra no tiene representatividad por DRE; sin embargo, esto no afecta la comparabilidad para los estratos a nivel nacional.

1.2. Pesos muestrales EM 2018

Los pesos ayudan a corregir la distribución de la muestra en los estratos y a expandir la información muestral a la población.

El peso es el inverso a la probabilidad de selección de un conglomerado (IE) en el interior de cada estrato. En el caso del muestreo proporcional al tamaño, se utilizan las probabilidades conjuntas de selección de la IE. La inversa de esta probabilidad vendría a ser el peso de la IE; es decir:

$$pIE_{ij} = \frac{1}{p_{ij}} \tag{1.1}$$

donde:

p_{ij} = probabilidad de selección conjunta de la i-ésima IE en el j-ésimo estrato

pIE_{ij} = peso de la i-ésima IE y en el j-ésimo estrato

[4]Para las ECE de 2.° grado de primaria, desde el año 2007, se diseñó una muestra de control que tuvo como propósito reforzar la estandarización de la aplicación de las pruebas. Esta muestra era probabilística, estratificada y por conglomerados, y tenía un alcance nacional y por estrato de gestión, área y característica, y DRE. Los resultados oficiales de la ECE para este grado se han reportado en base a la muestra de control, la cual es muy similar a lo reportado por todos los estudiantes evaluados.

Peso por sección. Este es el inverso de la probabilidad de selección de las secciones.

$$psec_i = \frac{secIE_i}{seceva_i} \tag{1.2}$$

donde:
$psec_i =$ peso por sección en la i-ésima IE
$secIE_i =$ total de secciones de la i-ésima IE
$seceva_i =$ secciones evaluadas en la i-ésima IE

Ajuste por estudiantes no evaluados en la sección. Esta corrección se realiza de manera separa para cada área evaluada (Lectura y Matemática).

$$a_{ki} = \frac{t_{ki}}{s_{ki}} \tag{1.3}$$

donde:
$a_{ki} =$ ajuste por estudiante no evaluado en la k-ésima sección y en la i-ésima IE
$t_{ki} =$ estudiantes que asisten a la k-ésima sección en la i-ésima IE
$s_{ki} =$ estudiantes evaluados en la k-ésima sección y en la i-ésima IE

Peso final. Se obtiene un peso para Lectura y otro para Matemática, el cual está dado por la siguiente fórmula:

$$\boldsymbol{pf_{kij} = pIE_{ij} \times psec_i \times a_{ki}} \tag{1.4}$$

donde:
$\boldsymbol{pf_{kij}} =$ peso final en la k-ésima sección, en la i-ésima IE y el j-ésimo estrato
$\boldsymbol{pIE_{ij}} =$ peso de la i-ésima IE en el j-ésimo estrato
$\boldsymbol{psec_i} =$ peso por sección en la i-ésima IE
$\boldsymbol{a_{ki}} =$ ajuste por estudiantes no evaluados en la k-ésima sección y en la i-ésima IE

1.3. Validación de la muestra 2018

Para validar la muestra, se procedió a recalcular los resultados de la ECE 2016 de 2.° grado de primaria usando solo las escuelas seleccionadas en la EM 2018 siguiendo el procedimiento en cuanto a la selección de dos secciones al azar cuando la IE tenía 3 o más. El objetivo de este ejercicio fue observar si la muestra

seleccionada para el 2018 reflejaba el resultado censal del año 2016[5]. A continuación, se presentan estos ejercicios de comparación de resultados en Lectura y Matemática acompañados del intervalo de confianza calculado para las estimaciones.

Tabla 1.2. ***Comparación de resultados censales de Lectura de 2016 con los resultados que se obtendría con la muestra de escuelas de 2018***

	Censo 2016 (%)			Muestra 2018 (%)		
	En inicio	En proceso	Satisfactorio	En inicio	En proceso	Satisfactorio
Nacional	5,7	46,2	48,1	5,4	45,9	48,7
Estatal	6,7	47,1	46,2	5,9	45,7	48,3
No estatal	3,0	44,1	52,9	4,1	46,3	49,6
Urbana	3,7	43,9	52,4	3,9	43,1	52,9
Rural	18,8	61,9	19,2	15,5	64,6	19,9

Tabla 1.3. ***Intervalo de confianza para los resultados de 2016 a partir de la muestra de escuelas de 2018***

	En inicio (%)		En proceso (%)		Satisfactorio (%)	
	LI	LS	LI	LS	LI	LS
Nacional	4,2	6,6	43,3	48,5	45,8	51,6
Estatal	4,4	7,5	42,4	49,0	44,4	52,3
No estatal	1,9	6,3	42,0	50,6	44,9	54,3
Urbana	2,7	5,1	40,3	46,0	49,7	56,2
Rural	11,2	19,9	59,3	69,8	14,4	25,4

Nota. LI = límite inferior, LS = límite superior.

Tabla 1.4. ***Comparación de resultados censales de Matemática de 2016 con los resultados que se obtendría con la muestra de escuelas de 2018***

	Censo 2016 (%)			Muestra 2018 (%)		
	En inicio	En proceso	Satisfactorio	En inicio	En proceso	Satisfactorio
Nacional	26,6	37,4	36,1	26,8	35,8	37,4
Estatal	23,5	37,1	39,4	22,0	35,6	42,4
No estatal	34,4	38,1	27,4	39,3	36,4	24,4
Urbana	23,7	37,8	38,4	24,8	36,1	39,1
Rural	45,5	34,2	20,2	39,9	34,2	26,0

[5]Cabe resaltar que no todas las escuelas seleccionadas en la EM 2018 en 2.° grado de primaria fueron evaluadas en la ECE 2016, esto debido a que la ECE solo evalúa a aquellas escuelas que tienen 5 o más estudiantes en el año a evaluar y algunas escuelas pueden participar o no en las distintas evaluaciones. El porcentaje de escuelas de la EM 2018 que también fue evaluado en la ECE 2016 fue de 89,5 %.

Tabla 1.5. ***Intervalo de confianza para los resultados de 2016 a partir de la muestra de escuelas de 2018***

	En inicio (%)		En proceso (%)		Satisfactorio (%)	
	LI	LS	LI	LS	LI	LS
Nacional	24,3	29,3	34,0	37,7	34,5	40,3
Estatal	18,9	25,0	33,4	37,9	38,6	46,2
No estatal	34,3	44,2	33,5	39,2	20,7	28,0
Urbana	21,9	27,8	34,1	38,1	35,7	42,5
Rural	32,8	46,9	29,3	39,1	19,5	32,4

Nota. LI = límite inferior, LS = límite superior.

En general, se observa que el resultado censal 2016 se encuentra dentro de los intervalos de confianza calculados a partir de las escuelas que participaron en la EM 2018 tanto en Lectura como en Matemática. Como ejemplo, se puede tomar el resultado en nivel Satisfactorio de la ECE 2016, 27,4 %, y contrastarlo con el resultado que se obtiene utilizando la muestra de la EM 2018, 24,4 % (tabla 1.4). En la tabla 1.5 se puede observar que el resultado de la ECE 2016 se encuentra dentro del intervalo de confianza calculado con la muestra del 2018 (20,7 % - 28,0 %), por lo que no presenta una diferencia estadísticamente significativa. Resultados similares se encuentran con el resto de niveles en Lectura y Matemática. Este ejercicio de validación permite asegurar que la muestra representa adecuadamente los resultados de la población y que no habría sesgo alguno.

Proceso de aplicación de las pruebas EM 2018

Capítulo 2

Capítulo 2

Proceso de aplicación de las pruebas EM 2018

La UMC se encarga de asegurar la implementación de controles de calidad que avalen la confiabilidad de los resultados obtenidos en las evaluaciones que están bajo su cargo a través de procedimientos explicitados en términos de referencia y manuales de procedimiento, y del monitoreo constante con personal capacitado a nivel nacional.

Además, los procesos de aplicación, sean ejecutados directamente por la UMC o a través de un operador logístico, se rigen bajo tres principios fundamentales que aseguran la confiabilidad de los resultados: estandarización, confidencialidad y probidad. Estos principios se vienen siguiendo en operativos censales, muestrales, pilotos, nacionales e internacionales, y en todos los tipos de evaluación de los que es responsable la UMC.

Para asegurar el cumplimiento de estos principios, se llevan a cabo controles de calidad durante todo el proceso de aplicación. Por un lado, estos controles de calidad se enfocan en el trabajo que realiza el aplicador, pues es la persona encargada de aplicar los cuadernillos de pruebas y del traslado de respuestas a la ficha óptica. En particular, uno de los aspectos en los que más se incide es en la labor que realiza el aplicador dentro del aula y en el manejo que hará de los instrumentos. El objetivo es asegurar que las acciones del aplicador estén completamente pauteadas y estandarizadas, a fin de reducir la interferencia de variables o componentes externos al proceso de aplicación en el aula.

Los controles de calidad implementados para la aplicación de la prueba consisten en una serie de procesos estandarizados para la capacitación y selección de personal. Estos procesos aseguran la captación del personal más idóneo y, al mismo tiempo, una transmisión eficiente de los procedimientos que se ejecutarán en el aula durante la aplicación. Uno de los primeros procesos consiste en la elaboración de un manual de procedimientos estandarizados. Todos los procedimientos de aplicación, manejo de instrumentos, casuísticas y ejemplos prácticos se encuentran descritos a detalle en este documento, el cual sirve como herramienta de consulta obligatoria y permanente para los aplicadores, sobretodo durante los días de aplicación.

Otro de los procesos orientados a asegurar la calidad de la aplicación es diseñar una capacitación estandarizada. El diseño de capacitación transmite los procedimientos descritos en el manual a los candidatos a aplicadores de manera clara y directa a través de una metodología sobria y de fácil replicación. Asimismo, el diseño dispone

de momentos en los que los candidatos a aplicadores se ejercitan en el manejo de los instrumentos y simulan la aplicación en aula siendo evaluados constantemente.

En la capacitación para la EM 2018 de 2.° grado de primaria, se preparó al aplicador enfatizando y reforzando los procedimientos para el manejo de la Ficha óptica de asistencia y respuesta (FOAR), la cual es el instrumento donde se trasladan las respuestas. Esto se realiza con ejercicios estandarizados y calificados haciendo uso de réplicas exactas del instrumento.

Finalmente, en la EM 2018, tal y como se hace en todos los operativos de la oficina, el candidato a aplicador debe demostrar los conocimientos aprendidos aprobando un examen al final de la capacitación. Los candidatos con mejores puntajes son los que son finalmente seleccionados como aplicadores.

Adicionalmente, en la EM 2018 se mantuvo la decisión de contratar a los aplicadores que evidenciaron un mejor desempeño en las aplicaciones de las ECE. Es decir, para la EM se renovó la contratación de los aplicadores que ejecutaron mejor los procedimientos de aplicación en los otros procesos de evaluación que iban finalizando.

Por otro lado, además de los controles de calidad que se implementan para el trabajo de los aplicadores en el aula, también se implementan controles de calidad que tienen que ver con el aseguramiento de las condiciones de aplicación y de la confidencialidad de los instrumentos hasta las fechas en que deben ser aplicados. Por eso, siguiendo la experiencia de operativos anteriores, se dispone que los instrumentos de aplicación estén salvaguardados en almacenes bajo la responsabilidad del operador logístico o del representante de la UMC en la sede de aplicación. Este almacenaje y resguardo se realiza siguiendo los lineamientos dictados por la oficina y bajo estricto monitoreo. Ello asegura la confidencialidad de todos los instrumentos a ser utilizados en la aplicación.

Las condiciones de aplicación también son estandarizadas a nivel nacional en todas las instituciones educativas en las que se lleva a cabo. La EM 2018 tuvo lugar los días 13 y 14 de noviembre en las IE de la muestra. Ambos días se evaluaron Lectura y Matemática, pero en distinto orden. El primer día, primero se evalúo Lectura; el segundo día, primero Matemática. Cada prueba duró un total de 45 minutos.

Con el fin de corroborar la estandarización de los procedimientos y el trabajo del aplicador, el director de la IE revisa algunas fichas de asistencia y respuestas (FOAR) al azar luego que el aplicador ha trasladado las respuestas de los estudiantes en ellas al interior de la IE. Este proceso de revisión sigue el siguiente procedimiento: 1) el director recibe la FOAR y escoge al azar, como mínimo, diez números de correlativos del estudiante; 2) el aplicador saca el paquete de cuadernillos de pruebas y busca los que correspondan a los correlativos seleccionados; 3) el aplicador lee en voz alta

el correlativo, apellidos y nombres del estudiante del cuadernillo de pruebas y pide al director que verifique estos datos en la FOAR; 4) se leen en voz alta las respuestas marcadas por el estudiante en el cuadernillo (número y letra); 5) el director verifica lo dictado en la FOAR en cada uno de los cuadernillos escogidos[6]; 6) una vez hecha la verificación, se guardan los cuadernillos en una bolsa de seguridad en presencia del director hasta su llegada al centro de captura de datos; y 7) el director llena una declaración jurada indicando si todos los procedimientos que se indican en ella se han cumplido.

Finalizada la aplicación en la IE, los aplicadores deben dirigirse inmediatamente hacia la sede operativa o al encuentro del representante de la oficina con todos los instrumentos de aplicación, los cuales son manejados siguiendo procedimientos que aseguran su confidencialidad después de su aplicación. Todos los controles de calidad que implementa la oficina están orientados a asegurar el cumplimiento estandarizado de estas condiciones de aplicación.

Por último, cuando todos los instrumentos aplicados retornan a Lima, el equipo de Análisis de la UMC se encarga de la detección de errores en la base de datos de 2.° grado de primaria. En la EM 2018, del universo total de respuestas de todos los estudiantes en todas las instituciones educativas participantes, se detectó un total de 364 errores en los que el aplicador olvidó marcar una opción en la ficha óptica. Esta cantidad representa solo el 0,04 % del total de alternativas codificadas, por lo que no afecta los resultados obtenidos.

Además, se verificó la existencia de inconsistencias en el llenado de la declaración jurada del director y la declaración jurada del aplicador en la FOAR. Específicamente, se corroboraron las secciones correspondientes a la verificación del traslado de respuestas. Para realizar ello, se utilizó la base de datos de la FOAR y se verificó si es que había consistencia entre el ítem 5.1 de la declaración jurada del director (*"verifiqué en la ficha óptica junto con el aplicador, el traslado de respuestas en una muestra de diez estudiantes como mínimo"*) y el ítem 9 de la declaración jurada del aplicador (*"cada día realicé la verificación del traslado de respuestas de los cuadernillos a la FOAR, mínimo 10, junto con el Director"*). Luego del análisis de todas las secciones, solo se encontraron 2 secciones en toda la muestra que no realizaron esta verificación de respuestas, lo que representa el 0,36 % de todas las secciones evaluadas. Este porcentaje no afecta los resultados.

[6] Si el director solicita revisar el traslado de respuestas de más estudiantes o de todos los estudiantes de la sección, el aplicador debe acceder a realizarlo.

Procesamiento de datos y análisis psicométrico

Capítulo 3

Capítulo 3

Procesamiento de datos y análisis psicométrico

La UMC ejecuta un protocolo que contempla los lineamientos establecidos por la *American Psychological Association* (APA), *la American Educational Research Association* (AERA) y el *National Council on Measurement in Education* (NCME). Ello permite garantizar la calidad de la información obtenida mediante la aplicación de sus evaluaciones. En este capítulo se detallan cada uno de estos procesos.

3.1. Gestión y depuración de datos

Este proceso tiene como finalidad convertir la información contenida en los documentos físicos (las fichas ópticas) en información digital (base de datos) consistente y confiable. Para ello, la UMC elabora especificaciones técnicas detallando los procedimientos que se requieren para un correcto procesamiento de la información. Este proceso de depuración puede dividirse en tres etapas principales: gestión física, captura de datos e imágenes y depuración de bases de datos.

Luego, se contrata a una empresa que cuente con experiencia en este rubro y que siga los procedimientos bajo la atenta supervisión de la UMC. Dicha empresa provee todo el personal necesario, el cual es evaluado, seleccionado y capacitado por los representantes de la UMC.

3.1.1. Gestión física

Este proceso tiene como objetivo asegurar que toda la documentación física a procesar esté completa y lista para la captura de datos. En este momento, se realizan los siguientes pasos:

- Recepción e inventario: La empresa de captura de datos recibe todos los documentos físicos y realiza un inventario uno a uno escaneando el código de barras impreso en cada documento.
- Clasificación: Se clasifican todos los documentos físicos según su tipo, los cuales pueden ser fichas ópticas de respuesta, cuestionarios u otros.
- Preparación: Consiste en agrupar los documentos físicos en lotes pequeños y quitar las grapas de estos para una mejor gestión en la siguiente etapa.
- Almacenamiento: Todos los documentos físicos son almacenados por la empresa de captura de datos de forma ordenada para poder acceder a estos en caso se requiera durante la siguiente etapa.

Para asegurar una correcta gestión física de los documentos, la empresa de captura de datos proporciona un local con el espacio suficiente para todas las etapas, incluyendo un espacio exclusivo para el almacenamiento de todos los documentos debido a la confidencialidad de los mismos.

3.1.2. Captura de datos

Este proceso tiene como objetivo digitalizar los documentos físicos y capturar los datos consignados en los mismos.

- Digitalización y captura de datos: Consiste en pasar por un escáner especial todos los documentos físicos ya preparados. El escáner utilizado es capaz de digitalizar el documento físico en una imagen con buena calidad, así como capturar los datos consignados en este utilizando tres tecnologías: OMR (Reconocimiento de marcas tipo burbujas), OCR (Reconocimiento de caracteres impresos) e ICR (Reconocimiento de caracteres escritos a mano).
- Control de calidad: Esta tarea se realiza cuando el software de reconocimiento de datos presenta un bajo nivel de confiabilidad de la data capturada. Estos casos son enviados a un grupo de personas dedicadas a revisar y asegurar que el dato capturado corresponda con el dato consignado en la imagen del documento.

3.1.3. Depuración de datos

En este proceso, se verifica que todos los campos de las bases de datos contengan respuestas que estén dentro de los márgenes permitidos y que exista coherencia entre los campos y la información adicional que se maneja en la evaluación. Para este proceso, la UMC elabora un manual del depurador, el cual contiene todas las revisiones a realizar en cada campo de todas las bases de datos. Además de este manual, los depuradores son capacitados para que realicen de la forma más eficiente posible dicho proceso. Durante esta etapa de depuración, algunas de las revisiones realizadas son las siguientes:

- Revisiones iniciales:
    - Correspondencia de campos según “diccionario de variables” y la frecuencia de cada campo
    - Correspondencia de registros según el reporte de inventario de instrumentos
    - Registros únicos en los campos de identificación del instrumento como el código de barras
    - Correspondencia entre código modular y correlativo Minedu (según padrón de IE)
    - Hojas faltantes en cada ficha óptica

- Revisiones principales (por tipo de documento):
    - Caracteres no válidos en los DNI de los estudiantes
    - Duplicados en los DNI de los estudiantes
    - Caracteres no válidos en los nombres y apellidos de los estudiantes
    - Duplicados en los apellidos y nombres de los estudiantes
    - Caracteres no válidos en el correlativo del estudiante
    - Duplicados en el correlativo de estudiante
    - Caracteres no válidos en el sexo del estudiante
    - Caracteres no válidos en la lengua materna del estudiante
    - Caracteres no válidos en la discapacidad del estudiante
    - Caracteres no válidos en el número de forma del cuadernillo
    - Caracteres no válidos en las respuestas del estudiante
    - Inconsistencia en las respuestas y la "asistencia" del estudiante
- Revisiones finales:
    - Repetir todas las revisiones iniciales
    - Contrastar con el listado de IE aplicadas proporcionado por el equipo encargado del trabajo campo de la UMC
    - Identificar valores fuera del rango permitido

3.2. Procesamiento psicométrico

A diferencia de las pruebas de Matemática y Lectura aplicadas en las ECE –donde se aplicó una única prueba para todos los estudiantes–, las pruebas de la EM 2018 de 2.° grado de primaria usaron una estructura con bloques rotados; es decir, los estudiantes se enfrentaron a distintos conjuntos de ítems, pero con un subconjunto de ellos en común. Este cambio no afecta la comparabilidad de los resultados entre años, ya que gracias al modelo psicométrico utilizado la medida de habilidad estimada de cada estudiante no depende del conjunto de ítems aplicados. Al contrario, esta nueva estructura ofrece algunas ventajas para la medición, como la posibilidad de contrarrestar la copia y la de tener una mayor cantidad de ítems para la estimación de las habilidades. Asimismo, se aplicaron pruebas diferenciadas para la población urbana y rural con el fin de mejorar la confiabilidad de las medidas al tener ítems con dificultades cercanas a la distribución de habilidad de la población.

En las tablas 3.1 y 3.2 se presenta el diseño de bloques de la prueba de Lectura; lo mismo para Matemática en las tablas 3.3 y 3.4.

Tabla 3.1. ***Matriz de bloques de la prueba de Lectura. Día 1***

Forma	Bloque	Ítems por cuadernillo
1	BC1	25
	B01	
	B02	
2	BC1	25
	B03	
	B04	
3	B05	25
	BC1	
	B06	
	B07	
4	B08	25
	BC1	
	B09	
	B02	

Tabla 3.2. ***Matriz de bloques de la prueba de Lectura. Día 2***

Forma	Bloque	Ítems por cuadernillo
5	BC2	25
	BC4	
	B10	
	B11	
6	BC2	25
	BC4	
	B12	
	B13	
7	B14	25
	BC2	
	BC3	
	B15	
	B13	
8	B16	25
	BC2	
	BC3	
	B17	
	B18	

Tabla 3.3. ***Matriz de bloques de la prueba de Matemática. Día 1***

Forma	Bloque	Ítems por cuadernillo
1	B03	23
	B11	
	B01	
2	B05	23
	B12	
	B01	
3	B03	23
	B04	
	B01	
4	B05	23
	B06	
	B01	

Tabla 3.4. ***Matriz de bloques de la prueba de Matemática. Día 2***

Forma	Bloque	Ítems por cuadernillo
5	B07	23
	B13	
	B02	
6	B09	23
	B14	
	B02	
7	B07	23
	B08	
	B02	
8	B09	23
	B10	
	B02	

Estos cambios se implementaron debido a que en mediciones anteriores se observó que, con el pasar de los años, las pruebas eran fáciles de responder por los estudiantes más hábiles. El nuevo diseño de pruebas permitió mejorar esta distribución (figuras 1 y 2 del anexo) y obtener mejores indicadores psicométricos que en los procesos anteriores (por ejemplo, la confiabilidad de las medidas).

En cuanto al modelo psicométrico utilizado, se aplica el *Rasch para ítems dicotómicos*, el cual permite estimar la *dificultad* de cada ítem o pregunta, así como la *habilidad* de los estudiantes evaluados, colocando ambas medidas en una misma métrica.

Una parte central del proceso de análisis psicométrico radica en verificar que los datos se ajusten al modelo Rasch. Para ello se utilizan los indicadores *infit* y *outfit*, los cuales reportan el nivel de adecuación de las respuestas en el centro y los

extremos de la distribución de habilidad según lo esperado por el modelo. Además, se emplean otros indicadores relevantes como la *correlación ítem-medida* (ptme) y la *tasa de acierto* (p) de cada ítem. De esta manera, el modelo Rasch permite establecer un control de calidad de los insumos utilizados en la evaluación de los estudiantes. Así, por ejemplo, en caso de detectarse ítems con desajuste (con patrones extraños de respuesta), estos no son utilizados en la estimación de las medidas de habilidad de los estudiantes.

3.2.1. Calibración de ítems

Las figuras 1 y 2 del anexo muestran una distribución de medidas de habilidad bastante simétrica con la de las medidas de dificultad de los ítems en ambas áreas. Esto quiere decir que existen ítems mejor distribuidos a lo largo de la escala de habilidad de los estudiantes.

Además, como se mencionó en el acápite anterior, los ítems de las pruebas presentan adecuados indicadores de ajuste y otros relevantes en el proceso psicométrico. A continuación, el detalle de ellos en las tablas 3.5 y 3.6:

Tabla 3.5. ***Dificultad y ajuste de los ítems al modelo Rasch en Lectura***

Bloque	Orden	Medida	Error	Infit	Outfit	ptme	p	Nivel
B01	15	1,85	0,031	0,91	0,86	0,51	0,58	2
B01	16	2,604	0,031	1,01	1,08	0,42	0,42	3
B01	17	2,565	0,031	0,96	1,01	0,46	0,43	3
B01	18	3,062	0,034	1,04	1,09	0,37	0,35	3
B01	19	2,678	0,033	1,18	1,25	0,26	0,43	3
B01	20	2,487	0,033	1,16	1,22	0,28	0,48	3
B02	21	0,676	0,037	0,85	0,7	0,5	0,8	2
B02	22	1,949	0,03	1,09	1,11	0,37	0,56	2
B02	23	2,509	0,03	1,25	1,38	0,23	0,44	3
B02	24	-	-	-	-	-	-	-
B02	25	2,589	0,031	1,14	1,25	0,32	0,42	3
B03	15	1,824	0,034	1,05	1,07	0,38	0,62	2
B03	16	1,531	0,032	0,97	0,9	0,47	0,65	2
B03	17	1,243	0,033	0,9	0,78	0,5	0,71	2
B03	18	2,428	0,033	1	1,01	0,43	0,49	3
B03	19	1,948	0,031	0,86	0,81	0,56	0,56	2
B03	20	2,284	0,03	1,02	1,01	0,44	0,49	3
B04	21	0,695	0,042	0,83	0,63	0,5	0,82	2
B04	22	-	-	-	-	-	-	-
B04	23	2,375	0,033	1,1	1,15	0,35	0,5	3
B04	24	2,403	0,033	0,99	1	0,44	0,5	3
B04	25	2,763	0,034	1,05	1,08	0,39	0,42	3
B05	1	-0,221	0,092	0,96	0,81	0,37	0,83	1
B06	17	1,844	0,072	1,07	1,08	0,36	0,42	2
B06	20	2,157	0,074	1,07	1,09	0,36	0,36	3

Bloque	Orden	Medida	Error	Infit	Outfit	ptme	p	Nivel
B07	21	0,907	0,073	0,89	0,83	0,5	0,63	2
B07	22	1,963	0,073	1,01	1	0,41	0,4	2
B07	23	1,701	0,071	0,96	0,94	0,46	0,46	2
B07	24	1,767	0,072	0,99	0,99	0,43	0,44	2
B07	25	1,693	0,071	1,05	1,06	0,38	0,46	2
B08	1	0,304	0,081	0,96	0,93	0,41	0,75	2
B09	17	1,757	0,072	0,92	0,89	0,51	0,45	2
B10	38	0,792	0,041	0,87	0,71	0,47	0,81	2
B10	39	1,311	0,036	0,94	0,85	0,45	0,73	2
B10	40	1,141	0,038	0,89	0,76	0,48	0,76	2
B10	41	1,122	0,038	0,91	0,78	0,46	0,76	2
B10	42	2,028	0,033	0,98	0,97	0,43	0,58	3
B10	43	2,514	0,033	1,09	1,15	0,34	0,47	3
B11	44	2,893	0,034	1,15	1,28	0,28	0,39	3
B11	45	2,391	0,03	1,08	1,18	0,37	0,46	3
B11	46	2,56	0,03	1,1	1,19	0,35	0,43	3
B11	47	2,644	0,031	0,92	0,94	0,49	0,41	3
B11	48	2,368	0,033	1,07	1,09	0,36	0,5	3
B12	38	2,594	0,031	1,1	1,27	0,35	0,42	3
B12	39	2,534	0,031	1,04	1,1	0,41	0,43	3
B12	40	1,637	0,031	1,01	1,01	0,43	0,63	2
B12	41	3,288	0,036	1,12	1,37	0,29	0,31	3
B12	42	2,958	0,034	1,05	1,17	0,38	0,38	3
B12	43	3,192	0,033	1,21	1,46	0,25	0,3	3
B13	44	1,716	0,031	1,01	1	0,43	0,61	2
B13	45	2,681	0,031	1,11	1,21	0,35	0,4	3
B13	46	2,377	0,03	1,02	1,05	0,43	0,47	3
B13	47	1,939	0,031	1,14	1,19	0,33	0,56	2
B13	48	1,728	0,031	0,98	0,99	0,45	0,61	2
B14	24	0,632	0,076	0,92	0,83	0,46	0,69	2
B15	41	1,213	0,071	1,01	0,97	0,41	0,57	2
B16	24	0,559	0,076	0,87	0,76	0,51	0,7	2
B17	39	0,91	0,073	1,07	1,03	0,37	0,63	2
B17	40	1,611	0,071	1,01	1,02	0,43	0,48	2
B17	41	1,537	0,071	1,11	1,1	0,35	0,49	2
B17	42	1,106	0,071	0,94	0,88	0,49	0,59	2
B17	43	1,209	0,071	0,95	0,93	0,48	0,57	2
B18	45	1,954	0,072	1,1	1,14	0,35	0,4	2
B18	48	2,352	0,075	1,27	1,41	0,19	0,32	3
BC1	1	-0,645	0,041	0,89	0,61	0,36	0,94	1
BC1	2	-0,535	0,039	0,87	0,51	0,39	0,93	1
BC1	3	-0,306	0,036	0,87	0,59	0,41	0,91	1
BC1	4	1,128	0,024	0,99	0,96	0,42	0,73	2
BC1	5	0,418	0,028	0,89	0,81	0,44	0,84	2
BC1	6	1,372	0,023	0,94	0,89	0,47	0,68	2
BC1	7	0,892	0,025	0,94	0,88	0,44	0,77	2
BC1	8	2,175	0,021	0,97	0,96	0,47	0,51	3

Bloque	Orden	Medida	Error	Infit	Outfit	ptme	p	Nivel
BC1	9	2,271	0,021	0,98	0,96	0,47	0,49	3
BC1	10	2,11	0,021	1,1	1,14	0,37	0,53	3
BC1	11	0,465	0,028	0,89	0,72	0,46	0,83	2
BC1	12	0,436	0,028	0,9	0,86	0,43	0,84	2
BC1	13	1,301	0,023	0,96	0,94	0,45	0,7	2
BC1	14	1,673	0,022	1,08	1,09	0,37	0,62	2
BC2	24	0,436	0,028	0,9	0,74	0,44	0,83	2
BC2	25	0,451	0,028	0,89	0,77	0,44	0,83	2
BC2	26	2,014	0,021	1,11	1,14	0,36	0,55	2
BC2	27	1,841	0,022	0,96	0,95	0,47	0,58	2
BC2	28	1,046	0,024	0,95	0,87	0,45	0,74	2
BC2	29	0,895	0,025	0,85	0,71	0,52	0,77	2
BC2	30	0,827	0,025	0,84	0,68	0,52	0,78	2
BC2	31	2,354	0,021	1,15	1,24	0,32	0,47	3
BC2	32	1,984	0,021	1,19	1,31	0,29	0,55	2
BC2	33	1,026	0,024	0,94	0,86	0,45	0,74	2
BC3	35	0,798	0,052	0,9	0,82	0,5	0,66	2
BC3	36	0,866	0,052	0,87	0,81	0,52	0,64	2
BC3	37	0,332	0,056	0,85	0,71	0,51	0,75	2
BC3	38	1,628	0,05	1,22	1,33	0,24	0,47	2
BC4	34	2,115	0,023	1,02	1,02	0,41	0,56	3
BC4	35	2,648	0,024	1,18	1,25	0,28	0,44	3
BC4	36	2,025	0,024	0,95	0,95	0,46	0,58	3
BC4	37	1,329	0,026	0,83	0,74	0,54	0,72	2

Tabla 3.6. ***Dificultad y ajuste de los ítems al modelo Rasch en Matemática***

Bloque	Orden	Medida	Error	Infit	Outfit	ptme	p	Nivel
B01	40	-0,28	0,022	1,01	1,06	0,39	0,63	1
B01	41	0,961	0,022	1,08	1,09	0,38	0,46	2
B01	42	2,412	0,024	1,12	1,22	0,35	0,27	3
B01	43	-0,489	0,022	0,95	0,92	0,44	0,66	1
B01	44	-0,22	0,022	0,96	0,98	0,43	0,63	1
B01	45	2,763	0,025	1,04	1,16	0,39	0,23	3
B01	46	1,39	0,022	0,91	0,88	0,52	0,4	2
B01	47	0,316	0,021	0,88	0,83	0,53	0,55	2
B01	48	1,475	0,022	0,85	0,79	0,57	0,39	2
B02	15	-1,621	0,025	0,92	0,84	0,41	0,79	1
B02	16	0,064	0,021	0,88	0,83	0,52	0,58	1
B02	17	2,15	0,023	0,96	0,94	0,48	0,3	3
B02	18	0,789	0,021	0,82	0,77	0,58	0,48	2
B02	19	0,439	0,021	1,05	1,05	0,4	0,53	2
B02	20	1,122	0,022	0,97	0,96	0,47	0,43	2
B02	21	1,48	0,022	0,96	0,94	0,48	0,39	2
B02	22	0,475	0,021	0,88	0,85	0,53	0,53	2
B02	23	0,857	0,021	1	1,01	0,44	0,47	2
B03	26	0,495	0,03	1,16	1,23	0,31	0,52	2

Bloque	Orden	Medida	Error	Infit	Outfit	ptme	p	Nivel
B03	27	-	-	-	-	-	-	-
B03	28	1,066	0,031	1	1,03	0,45	0,44	2
B03	29	0,252	0,03	0,94	0,92	0,47	0,56	2
B03	30	1,468	0,031	1,2	1,29	0,3	0,39	2
B03	31	1,4	0,031	1,17	1,27	0,32	0,4	2
B03	32	1,07	0,031	1,08	1,12	0,38	0,44	2
B03	33	1,304	0,031	1,24	1,32	0,27	0,41	2
B04	34	0,708	0,073	0,87	0,87	0,53	0,39	2
B04	35	0,296	0,072	0,98	1,08	0,43	0,45	2
B04	36	-1,044	0,073	1,15	1,48	0,23	0,64	1
B04	37	-1,114	0,073	0,9	0,8	0,46	0,65	1
B04	38	-0,67	0,071	0,95	0,9	0,43	0,59	1
B04	39	-0,028	0,071	1,17	1,16	0,28	0,5	1
B05	26	-0,317	0,031	1,08	1,15	0,34	0,64	1
B05	27	0,673	0,03	1,04	1,04	0,41	0,5	2
B05	28	0,631	0,03	1,2	1,26	0,28	0,51	2
B05	29	1,612	0,031	0,97	0,98	0,47	0,37	2
B05	30	0,361	0,03	0,9	0,85	0,51	0,55	2
B05	31	1,514	0,031	1,1	1,13	0,37	0,38	2
B05	32	1,239	0,031	1,06	1,07	0,4	0,42	2
B05	33	0,688	0,03	1,13	1,2	0,33	0,5	2
B06	34	-2,081	0,082	0,95	0,9	0,38	0,78	1
B06	35	-0,37	0,071	0,85	0,81	0,53	0,55	1
B06	36	-0,42	0,071	0,84	0,78	0,55	0,56	1
B06	37	-0,015	0,071	0,87	0,83	0,53	0,5	1
B06	38	0,211	0,071	0,95	0,94	0,46	0,46	2
B06	39	0,314	0,071	0,92	0,91	0,49	0,45	2
B07	1	0,818	0,03	0,89	0,87	0,53	0,48	2
B07	2	-0,019	0,031	0,9	0,88	0,5	0,59	1
B07	3	2,226	0,034	0,98	1,03	0,46	0,29	3
B07	4	0,64	0,03	0,85	0,8	0,56	0,5	2
B07	5	0,953	0,031	0,8	0,76	0,6	0,46	2
B07	6	-	-	-	-	-	-	-
B07	7	0,624	0,03	0,9	0,89	0,52	0,5	2
B07	8	1,657	0,032	0,91	0,91	0,52	0,36	2
B08	9	-3,47	0,108	1,01	1,19	0,25	0,89	1
B08	10	-1,542	0,076	0,99	0,95	0,36	0,71	1
B08	11	0,906	0,075	0,8	0,77	0,59	0,36	2
B08	12	-0,06	0,071	0,94	0,9	0,47	0,5	1
B08	13	0,429	0,072	0,85	0,79	0,55	0,43	2
B08	14	1,662	0,081	0,87	0,89	0,53	0,27	2
B09	1	-0,56	0,031	0,84	0,74	0,53	0,67	1
B09	2	0,144	0,03	1,1	1,16	0,34	0,58	2
B09	3	0,956	0,03	0,97	0,96	0,47	0,46	2
B09	4	0,832	0,03	0,86	0,83	0,55	0,48	2
B09	5	-	-	-	-	-	-	-
B09	6	0,308	0,03	1,02	1,01	0,42	0,55	2

Bloque	Orden	Medida	Error	Infit	Outfit	ptme	p	Nivel
B09	7	1,881	0,032	1,28	1,44	0,22	0,34	3
B09	8	1,03	0,03	1,05	1,07	0,41	0,45	2
B10	9	2,087	0,084	0,97	1,07	0,41	0,22	3
B10	10	-1,559	0,075	0,98	0,95	0,39	0,71	1
B10	11	-0,886	0,071	0,98	1,03	0,4	0,62	1
B10	12	-0,12	0,07	0,92	0,9	0,48	0,51	1
B10	13	0,194	0,07	0,91	0,93	0,49	0,46	2
B10	14	1,023	0,074	0,96	0,94	0,46	0,35	2
B11	34	1,641	0,034	1,05	1,07	0,41	0,38	2
B11	35	2,184	0,036	1,07	1,15	0,39	0,31	3
B11	36	2,031	0,036	0,85	0,8	0,57	0,33	3
B11	37	2,033	0,036	1,08	1,15	0,38	0,33	3
B11	38	2,53	0,038	0,99	1	0,45	0,27	3
B11	39	2,052	0,036	0,98	1,01	0,46	0,33	3
B12	34	2,012	0,035	1,02	1,02	0,43	0,34	3
B12	35	1,719	0,034	1,17	1,24	0,31	0,38	2
B12	36	1,978	0,035	1,08	1,1	0,38	0,34	3
B12	37	1,987	0,035	0,93	0,91	0,5	0,34	3
B12	38	3,629	0,044	1,07	1,21	0,33	0,17	3
B12	39	2,304	0,036	1,1	1,17	0,36	0,3	3
B13	9	2,99	0,04	1,25	1,39	0,25	0,22	3
B13	10	2,072	0,036	0,97	0,94	0,48	0,33	3
B13	11	1,743	0,035	1,16	1,22	0,33	0,37	2
B13	12	2,353	0,037	1,16	1,3	0,32	0,29	3
B13	13	1,832	0,035	1,14	1,17	0,35	0,36	2
B13	14	2,758	0,039	1,11	1,3	0,35	0,25	3
B14	9	2,692	0,038	1,21	1,46	0,25	0,26	3
B14	10	2,218	0,036	1,09	1,2	0,36	0,31	3
B14	11	-	-	-	-	-	-	-
B14	12	2,505	0,037	0,97	1	0,45	0,28	3
B14	13	0,794	0,033	0,86	0,81	0,55	0,51	2
B14	14	2,027	0,035	0,93	0,91	0,5	0,34	3

3.2.2. Evidencias de validez vinculadas al contenido

Los resultados de las pruebas de Lectura y Matemática de 2.° grado de primaria de la EM 2018 son comparables con los de la ECE 2016 puesto que las pruebas de las que se derivan tienen una alta correspondencia. Además de la evidencia estadística, esto se sostiene en la similitud entre las tablas de especificaciones usadas para diseñar ambas pruebas.

En Lectura, tal como se aprecia en la tabla 3.7, tanto en la ECE 2016 como en la EM 2018, existe un predominio de ítems que evalúan la capacidad “Infiere e interpreta

información del texto"; en segundo lugar, se encuentran los que miden la capacidad "Obtiene información del texto escrito" y; en tercer lugar, los que miden la capacidad "Reflexiona y evalúa la forma, el contenido y el contexto del texto".

En la EM 2018, se incluyó la capacidad "Lee oraciones", la cual fue evaluada hasta la ECE 2014. La decisión de incluir nuevamente esta capacidad se basa en la intención de medir con mayor precisión a los estudiantes con menor desarrollo de la competencia lectora. Por esa misma razón, hubo un pequeño incremento en la EM 2018 de la proporción de ítems de la capacidad "Obtiene información del texto escrito".

Respecto a la capacidad "Reflexiona y evalúa la forma, el contenido y el contexto del texto", también hubo un pequeño aumento en la proporción de ítems (de 10 % en la ECE 2016 a 14,6 % en la EM 2018). Este aumento se justifica en el necesario alineamiento de las pruebas al currículo. En tal sentido, era necesario representar esta capacidad con una mayor cantidad de ítems, pues constituye una capacidad fundamental para el desarrollo de la competencia lectora según el Currículo Nacional 2016 y los documentos que le precedieron. Esto además permitió generar una mayor dispersión en la escala de dificultad de los ítems para medir de forma más precisa a la población.

Los cambios descritos en los dos párrafos anteriores influyen en la disminución de la proporción de ítems de la capacidad "Infiere e interpreta información del texto" en la EM 2018. No obstante, cabe destacar que estos cambios son comunes de un año a otro y que se encuentran dentro de márgenes razonables que garantizan la comparabilidad de las pruebas.

Tabla 3.7. ***Lectura ECE 2016 y EM 2018. Distribución de ítems por capacidad en la prueba***

Capacidad	2016 (%)	2018 (%)
Lee oraciones	-	4,2
Obtiene información del texto escrito	30,0	33,3
Infiere e interpreta información del texto	60,0	47,9
Reflexiona y evalúa la forma, el contenido y el contexto del texto	10,0	14,6
Total	100	100

En el caso de Matemática, en ambas pruebas se evaluó la competencia "Resuelve problemas de cantidad", la cual involucra dos aspectos fundamentales: a) la construcción del significado y uso del número y del sistema de numeración decimal; y b) la construcción del significado y uso de las operaciones. Asimismo, se han evaluado todas las capacidades propuestas para la competencia. Aunque estas

capacidades son denominadas de formas diferentes en los currículos, estas en realidad se corresponden, tal como se observa en la siguiente tabla.

Tabla 3.8. ***Correspondencia de capacidades evaluadas en Matemática***

Capacidades ECE 2016	Capacidades EM 2018
Matematiza situaciones	Traduce cantidades a expresiones numéricas
Comunica y representa ideas matemáticas	Comunica su comprensión sobre los números y las operaciones
Elabora y usa estrategias	Usa estrategias y procedimientos de estimación y cálculo
Razona y argumenta generando ideas matemáticas	Argumenta afirmaciones sobre las estimaciones numéricas y las operaciones

En las tablas 3.9 y 3.10, se puede observar que la distribución de ítems (peso) por capacidad se mantiene en ambas pruebas. Además, se puede apreciar que tanto en la ECE 2016 como en la EM 2018, existe un predominio de ítems de la capacidad referida al modelamiento matemático (Matematiza situaciones / Traduce cantidades a expresiones numéricas); en segundo lugar, se encuentran los ítems de la capacidad referida a la comunicación y representación matemática (Comunica y representa ideas matemáticas / Comunica su comprensión sobre los números y las operaciones); en tercer lugar, los referidos al uso de diversas estrategias (Elabora y usa estrategias / Usa estrategias y procedimientos de estimación y cálculo); y, en cuarto lugar, los de la capacidad referida al razonamiento y argumentación matemáticos (Razona y argumenta generando ideas matemáticas / Argumenta afirmaciones sobre las estimaciones numéricas y las operaciones).

Tabla 3.9. ***Matemática ECE 2016. Distribución de ítems por capacidad en la prueba***

Capacidad	%
Matematiza situaciones	67,4
Comunica y representa ideas matemáticas	17,4
Elabora y usa estrategias	8,7
Razona y argumenta generando ideas matemáticas	6,5
Total	100

Tabla 3.10. ***Matemática EM 2018. Distribución de ítems por capacidad en la prueba***

Capacidad	%
Traduce cantidades a expresiones numéricas	62,2
Comunica su comprensión sobre los números y las operaciones	18,4
Usa estrategias y procedimientos de estimación y cálculo	8,2
Argumenta afirmaciones sobre las estimaciones numéricas y las operaciones	11,2
Total	100

Los ligeros cambios introducidos tuvieron la finalidad de enfatizar el alineamiento curricular de la prueba al enfoque del área (enfoque de resolución de problemas) y fortalecer el porcentaje de ítems de la capacidad referida al razonamiento y argumentación, que es una capacidad fundamental para el desarrollo de la competencia matemática. Otra finalidad de estos cambios fue generar una mayor dispersión de los ítems en la escala de dificultad, lo cual permite medir un rango más amplio de habilidad de los estudiantes.

3.2.3. Evidencias de validez vinculadas a la estructura interna

Uno de los supuestos importantes de los modelos Rasch es que los ítems a calibrar constituyen un conjunto unidimensional. Es decir, se parte del supuesto de que existe un rasgo latente predominante que explica las relaciones entre los ítems utilizados. En el caso de las pruebas aplicadas, no se encontraron evidencias serias que atenten contra este supuesto, lo cual aporta evidencias de validez referidas a la estructura interna de las pruebas.

Tabla 3.11. ***Análisis de unidimensionalidad de las medidas derivadas de la aplicación de las pruebas de Lectura y Matemática en la EM 2018***

Prueba	Varianza de la dimensión principal (%)	Varianza modelada (%)	Primer autovalor	% del primer autovalor
Lectura	27,5	27,9	2,18	1,7
Matemática	25,9	25,6	2,09	1,6

3.2.4. Confiabilidad y consistencia de la clasificación

El coeficiente utilizado es análogo al de consistencia interna alpha de Cronbach, pero produce mejores estimaciones, pues los valores numéricos son lineales si los datos se ajustan al modelo Rasch aplicado. Además, utiliza la varianza de error promedio de la muestra en lugar de la varianza de error de una persona promedio (Schumacker, 2007). El coeficiente sirve para indicar la capacidad de las medidas de un test para diferenciar las cantidades de rasgo latente que poseen los evaluados (Wrigth y Masters, 1982). En ese sentido, indica la replicabilidad del

ordenamiento de las personas según su medida de habilidad si se les da otro conjunto de ítems que miden el mismo constructo (Bond y Fox, 2015).

Un índice menor a 0,50 indica que las diferencias entre las medidas son producidas principalmente por el error de medición (Fisher, 1992). Sobre los valores mínimos aceptables de los coeficientes de confiabilidad, Charter (2003) ha realizado una revisión de numerosas investigaciones que proponen diferentes niveles mínimos. En ese estudio, encontró bastante variabilidad; asimismo, observó valores propuestos con los diversos métodos para obtener la confiabilidad, que oscilan entre 0,60 y 0,95. A pesar de esta gran variabilidad, un estándar mínimo aceptable que aparece con frecuencia en la literatura es el de 0,70, señalado por Nunnally y Bernstein (1995).

Estos coeficientes también pueden ser expresados como índices de separación de personas, que se refieren a la dispersión de los datos medidos como el número de errores estándar que separan a las personas (Schumacker, 2007). El índice de separación de personas (Gp) representa la variabilidad ajustada de las personas dividida entre el error estándar de medición promedio.

Tabla 3.12. ***Confiabilidad de las medidas de Lectura y Matemática en la EM 2018***

Prueba	Rp	Gp
Lectura	0,87	2,64
Matemática	0,88	2,69

Considerando el valor de Rp de las pruebas aplicadas en la EM 2018, se puede apreciar que la varianza de error es de 13 % en la prueba de Lectura y 12 % en la de Matemática, por lo cual es posible afirmar que las medidas derivadas de aplicar dichas pruebas poseen adecuadas evidencias de confiabilidad. Incluso, estos indicadores son más altos que los obtenidos en las pruebas de la ECE 2016.

Otros indicadores importantes para el análisis son los de precisión y consistencia de clasificación. La consistencia de la clasificación es el grado de acuerdo en dos administraciones independientes o paralelas de un instrumento de medición. Se espera que el alumno sea clasificado en la misma categoría al repetirse la evaluación. Por otro lado, la precisión de la clasificación implica el grado en el cual la clasificación observada coincide con la clasificación verdadera (Kim, Choi, Um y Kim, 2006).

En la práctica, es difícil de lograrlo, por ello, se han propuesto métodos para estimarla a partir de una sola aplicación. En la EM 2018, se utilizó el método de Rudner, que es una aproximación de tipo individual, pues calcula la consistencia de clasificación para cada persona y luego las promedia (Lee, 2010). Además, asume que los errores de estimación se distribuyen normalmente. A medida que aumenta el número de ítems,

dadas las propiedades del estimador $\hat{\theta}$ basado en un método de máxima verosimilitud, este supuesto es más plausible.

Tabla 3.13. ***Indicadores del análisis de precisión y consistencia de la clasificación a los niveles de logro en Lectura y Matemática en la EM 2018***

	Lectura		Matemática	
	Precisión	Consistencia	Precisión	Consistencia
Nacional	0,87	0,81	0,90	0,86
Hombres	0,86	0,80	0,90	0,86
Mujeres	0,87	0,82	0,90	0,86
Urbano	0,87	0,81	0,90	0,85
Rural	0,85	0,79	0,94	0,91
Estatal	0,86	0,81	0,90	0,86
No estatal	0,87	0,82	0,90	0,85

En general, la EM 2018 mantiene buenos indicadores de precisión y consistencia de la clasificación de los niveles de logro.

3.2.5. Equiparación de medidas

Una vez finalizada la estimación de la dificultad de los ítems y la habilidad de los evaluados, se procede con el proceso de equiparación. Es decir, mediante el empleo de ítems en común entre distintas pruebas que miden un mismo atributo, es posible estimar dificultades y habilidades en función de estimados obtenidos en una evaluación previa. Así, es posible hacer comparables los estimados de las dificultades de ítems aplicados en distintos momentos y en diferentes muestras o poblaciones de estudiantes.

La prueba del 2016 y la del 2018 tienen un grupo de ítems en común. Se encontró que en el 2018, en su mayoría, los ítems se han hecho más difíciles; es decir, las tasas de acierto (p) han disminuido en puntos porcentuales (Tablas 3.14 y 3.15). El mayor cambio se puede observar en uno de los ítems de la prueba de Matemática, el cual fue respondido por el 77 % de estudiantes en el 2016 pero, en el 2018, solo fue respondido por el 42 % de estudiantes de 2.° grado de primaria.

Tabla 3.14. ***Comparación de tasas de acierto de los ítems en común 2018 – 2016 en Lectura.***

Correlativo	p 2016	p 2018	2018-2016
1	0,58	0,49	-0,09
2	0,55	0,47	-0,08
3	0,63	0,55	-0,08
4	0,66	0,58	-0,08
5	0,58	0,51	-0,07
6	0,79	0,74	-0,05
7	0,58	0,53	-0,05
8	0,59	0,55	-0,04
9	0,87	0,83	-0,04
10	0,77	0,74	-0,03
11	0,93	0,91	-0,02
12	0,85	0,83	-0,02
13	0,79	0,77	-0,02
14	0,74	0,73	-0,01
15	0,79	0,78	-0,01
16	0,94	0,94	0,00
17	0,77	0,77	0,00
18	0,81	0,84	0,03

Tabla 3.15. ***Comparación de tasas de acierto de los ítems en común 2018 – 2016 en Matemática.***

Correlativo	p 2016	p 2018	2018-2016
1	0,77	0,42	-0,34
2	0,84	0,51	-0,33
3	0,68	0,35	-0,33
4	0,85	0,56	-0,30
5	0,68	0,39	-0,30
6	0,85	0,55	-0,30
7	0,65	0,37	-0,28
8	0,72	0,46	-0,26
9	0,50	0,23	-0,26
10	0,67	0,46	-0,21
11	0,61	0,40	-0,21
12	0,77	0,55	-0,21
13	0,72	0,53	-0,20
14	0,70	0,50	-0,19
15	0,72	0,53	-0,19
16	0,65	0,47	-0,18
17	0,79	0,63	-0,16
18	0,55	0,39	-0,16
19	0,82	0,66	-0,16
20	0,47	0,30	-0,16
21	0,64	0,50	-0,14
22	0,71	0,58	-0,13
23	0,59	0,46	-0,13
24	0,69	0,59	-0,09
25	0,43	0,34	-0,09
26	0,86	0,79	-0,07
27	0,71	0,67	-0,03

Se realizaron tres ejercicios de equiparación. El primero es el que se realiza en todos los procesos de evaluación; el segundo se trata de una calibración concurrente (calibración 2016 junto a 2018, en una sola base); y el tercero, consiste en aplicar la misma metodología del segundo ejercicio de equiparación, pero considerando solo las escuelas comunes (evaluadas en ambos años). Además, se sospechaba que la falta de ítems difíciles en la ECE 2016 había inflado los resultados de ese año, por lo que se eliminaron los ítems más difíciles del año 2018, para volver a realizar la calibración. En los cuatro escenarios se obtuvo una distribución similar de estudiantes por niveles de logro en 2.° grado de primaria.

A continuación, se presentan las tablas con los resultados de la equiparación por ítems comunes de las pruebas utilizadas en la EM 2018.

Tabla 3.16. ***Análisis DIF de las medidas de Lectura usadas en la equiparación 2016 con 2018***

Orden	2016		2018		Trans	DIF	
	Medida	Error	Medida	Error	Medida	Medida	Error
1	-0,708	0,010	-2,543	0,041	-0,644	-0,064	0,042
2	-0,448	0,009	-2,167	0,036	-0,305	-0,143	0,037
3	1,263	0,006	-0,579	0,024	1,128	0,135	0,025
4	0,437	0,007	-1,329	0,028	0,452	-0,015	0,029
5	2,084	0,005	0,402	0,021	2,014	0,070	0,022
6	0,948	0,006	-0,670	0,024	1,046	-0,098	0,025
7	0,920	0,006	-0,837	0,025	0,896	0,024	0,026
8	0,927	0,006	-0,912	0,025	0,828	0,099	0,026
9	2,327	0,005	0,778	0,021	2,353	-0,026	0,022
10	1,878	0,005	0,369	0,021	1,984	-0,106	0,022
11	1,072	0,006	-0,692	0,024	1,026	0,046	0,025
12	1,739	0,005	0,210	0,022	1,841	-0,102	0,023
13	1,059	0,006	-0,840	0,025	0,893	0,166	0,026
14	0,241	0,007	-1,346	0,028	0,436	-0,195	0,029
15	2,140	0,005	0,580	0,021	2,175	-0,035	0,022
16	0,768	0,006	-1,365	0,028	0,419	0,349	0,029
17	2,139	0,005	0,508	0,021	2,110	0,029	0,022
18	2,137	0,005	0,687	0,021	2,271	-0,134	0,022

Tabla 3.17. ***Análisis DIF de las medidas de Matemática usadas en la equiparación 2016 con 2018***

Orden	2016		2018		Trans	DIF	
	Medida	Error	Medida	Error	Medida	Medida	Error
1	0,656	0,006	-0,190	0,030	0,682	-0,026	0,031
2	-0,614	0,007	-0,940	0,071	-0,195	-0,419	0,071
3	0,838	0,006	0,046	0,031	0,957	-0,119	0,032
4	0,734	0,006	0,421	0,022	1,396	-0,662	0,023
5	-0,049	0,006	-0,796	0,022	-0,026	-0,023	0,023
6	0,596	0,006	-1,040	0,031	-0,312	0,908	0,032
7	-0,716	0,007	-1,802	0,025	-1,202	0,486	0,026
8	0,726	0,006	-0,652	0,031	0,142	0,584	0,032
9	1,006	0,006	-0,179	0,030	0,694	0,312	0,031
10	1,613	0,006	0,424	0,022	1,399	0,214	0,023
11	-0,542	0,007	-0,904	0,071	-0,153	-0,389	0,071
12	-0,488	0,007	-0,724	0,070	0,058	-0,546	0,070
13	0,793	0,006	0,096	0,074	1,016	-0,223	0,074
14	1,205	0,006	0,360	0,022	1,324	-0,119	0,023
15	2,343	0,006	0,806	0,035	1,845	0,498	0,036
16	0,158	0,006	-0,411	0,021	0,423	-0,265	0,022
17	-0,317	0,007	-0,989	0,022	-0,252	-0,065	0,023
18	0,504	0,006	-0,499	0,070	0,321	0,183	0,070
19	0,552	0,006	-0,592	0,021	0,212	0,340	0,022
20	0,470	0,006	-0,297	0,021	0,557	-0,087	0,022

Orden	2016		2018		Trans	DIF	
	Medida	Error	Medida	Error	Medida	Medida	Error
21	0,503	0,006	-0,323	0,021	0,526	-0,023	0,022
22	0,963	0,006	-0,023	0,021	0,877	0,086	0,022
23	2,101	0,006	0,905	0,023	1,961	0,140	0,024
24	0,157	0,006	-0,330	0,072	0,518	-0,361	0,072
25	1,341	0,006	0,052	0,022	0,964	0,377	0,023
26	1,255	0,007	0,519	0,031	1,510	-0,255	0,032
27	1,930	0,006	1,345	0,025	2,475	-0,545	0,026

3.3. Análisis adicionales

Se realizaron una serie de análisis que ayudasen a identificar comportamientos irregulares en los datos recolectados para la EM 2018. Dichos análisis, en conjunto se conocen como análisis forense de datos (Simon, 2012) y pueden ser aplicados desde diferentes perspectivas.

En general, todos los análisis realizados apuntan a que no existen evidencias suficientes para afirmar que exista un problema masivo en los datos recolectados. Con ello, se descartan problemas importantes en cuanto a los psicométricos de las pruebas y su proceso de aplicación.

3.3.1. Análisis de respuestas en blanco

Si bien los ítems que son dejados en blanco por los estudiantes que se enfrentan a las pruebas de la EM 2018 se califican con 0 puntos, se hizo el ejercicio de considerarlos como ítems no aplicados. Con esa recodificación se volvieron a calibrar los ítems y realizar todo el proceso de equiparación y se encontró que los resultados casi no cambiaron.

3.3.2. Desajuste de personas según el modelo psicométrico (respuestas inesperadas)

Este análisis trata de evaluar la discrepancia entre el patrón de respuestas observado de una persona y aquel esperado según un modelo psicométrico (Meijer y Sijtsma, 2001).

Si bien existen muchos índices para realizar este tipo de análisis, Karabatsos (2003) comparó el funcionamiento de 36 índices de desajuste en el contexto de un modelo Rasch (modelo psicométrico que se utilizó) destacando el funcionamiento del índice HT (Sijtsma, 1988; Sijtsma y Meijer, 1992). En este caso, se han utilizado los indicadores de ajuste propios del modelo Rasch, infit y outfit (Bond y Fox, 2015), ya que en estudios realizados desde la UMC han arrojado resultados similares a los del HT.

Al analizar los datos de la EM 2018, se encontró que en Lectura casi el 6 % de estudiantes evaluados presentaban un desajuste en su patrón, mientras que en

Matemática dicho porcentaje fue de casi 4 %. Estos resultados no son muy elevados y además son parecidos a los encontrados en la ECE 2016.

3.3.3. Análisis de cambios inusuales en la medida promedio y porcentaje de estudiantes en el nivel Satisfactorio

Uno de los análisis realizados implicó detectar cambios inusuales en la medida promedio de una IE entre años consecutivos. Para ello, se utilizó la g de Hedges (Kline, 2004), la cual fue ajustada según el cambio observado a nivel de agregado de una forma similar a como lo proponen Gaertner y McBride (2017). Adicionalmente, se estimó para cada IE el tamaño del efecto del cambio relativo en la proporción de estudiantes en el nivel Satisfactorio usando la h de Cohen (Cohen, 1992).

Los resultados en Lectura mostraron que casi el 14 % de las IE evaluadas presentan un cambio grande en la medida promedio en comparación con el cambio observado a nivel nacional. Este porcentaje es prácticamente idéntico al observado cuando se analizan los datos de la ECE 2016 comparados con la ECE 2015. En cuanto a Matemática, casi el 13 % de IE presentan este cambio grande. En este caso, el porcentaje disminuye con respecto a la comparación efectuada entre los años 2016 y 2015, donde se observa un porcentaje de casi 18 %.

En cuanto al cambio en términos del porcentaje de estudiantes en el nivel Satisfactorio, en Lectura, 9 % de las IE presentan un cambio grande y un 11 % los presentan en Matemática. Dichos resultados son idénticos a los observados en la comparación entre los años 2015 y 2016.

3.3.4. Predicción de un área sobre la base de la otra

Se construyó un modelo de regresión a nivel de agregado (en este caso, secciones) para predecir el valor promedio en la medida de un constructo a partir de la medida promedio del otro constructo. Es decir, se construyó primero un modelo para predecir la medida promedio en Lectura a partir de la de Matemática y luego otro para realizar la predicción en sentido inverso (Matemática a partir de Lectura).

Dado que la relación entre la medida promedio en Lectura y Matemática no es perfecta, existirán discrepancias entre los valores observados y predichos (residuos). Por ello, el objetivo de este tipo de análisis implica detectar aquellas secciones que tengan una discrepancia muy grande (en este caso se trabajó con un valor $> |3|$ en el residuo estandarizado) entre su medida promedio observada y predicha.

En ninguno de los dos modelos analizados se encontraron problemas serios, pues cerca del 1 % de las secciones analizadas presentaron una discrepancia importante

en los resultados en ambas áreas; resultado que es similar al encontrado cuando se analizan los datos de la ECE 2016.

3.3.5. Diferencias según día de aplicación

Dado que las pruebas son aplicadas durante dos días, y que ambos cuadernillos son elaborados a partir de la misma tabla de especificaciones, estas pueden ser consideradas equivalentes. Ello facilita que, a partir de una propiedad de los modelos Rasch conocida como invarianza (de Ayala, 2009), podamos esperar que el resultado de un estudiante en el día 1 no difiera dentro de los márgenes de error de su resultado en el día 2. Por ello, se analiza el porcentaje de estudiantes que presenta diferencias estadísticamente significativas en sus resultados según el día de aplicación.

En el caso de Lectura, se encontró que cerca del 9 % presentó diferencias estadísticamente significativas según el día de aplicación, mientras que en Matemática, dicho porcentaje fue de 11 %. Estos resultados no difieren demasiado de lo encontrado en la ECE 2016, donde se aprecia que el porcentaje de estudiantes con diferencias estadísticamente significativas es de 10 % en Lectura y 7 % en Matemática.

Referencias

Referencias

American Educational Research Association, American Psychological Association, y National Council of Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bond, T. G. y Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3.ª ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Charter, R. A. (2003). A breakdown of reliability coefficients by test type and reliability method, and the clinical implications of low reliability. *The Journal of General Psychology, 130*(3), 290-304.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.

de Ayala, R. J. (2009). *The theory and practice of Item Response Theory.* Nueva York: Guilford Press.

Fisher, W. (1992). Reliability statistics. En J. M. Linacre (Ed.), *Rasch Measurement Transactions part 2, 1996* (p. 238). Chicago: MESA Press.

Gaertner, M., y McBride, Y. (2017). Detecting unexpected changes in pass rates. A comparison of two statistical approaches. En G. J. Cizek y J. A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (p. 262-279). New York, NY: Routledge.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*(4), 277-298. https://doi.org/10.1207/S15324818AME1604_2

Kim, D., Choi, S., Um, K., y Kim, J. (2006, abril). *A comparison of methods for estimating classification consistency.* Trabajo presentado en el Annual Meeting of the National Council on Education in Measurement, San Francisco, CA.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Lee, W. (2010). Classification consistency and accuracy for complex assessments using Item Response Theory. *Journal of Educational Measurement, 47*(1).

Meijer, R. R., y Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*(2), 107-135. https://doi.org/10.1177/01466210122031957

Nunnally, J., y Bernstein, I. (1995). *Teoría psicométrica* (3a ed.). México: McGraw-Hill.

Schumacker, R. (2007). Reliability: A Rasch perspective. *Educational and Psychological Measurement, 67*(3), 394-409.

Sijtsma, K. (1988). *Contributions to Mokken's non-parametric Item Response Theory* (Tesis doctoral no publicada).

Sijtsma, K., y Meijer, R. R. (1992). A Method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement, 16*(2), 149-157. https://doi.org/10.1177/014662169201600204

Simon, M. (2012, octubre). *Local outlier detection in data forensics: Data mining approach to flag unusual schools.* Trabajo presentado en The 2nd Annual Conference on Statistical Detection of Potential Test Fraud, Madison, WI.

Wrigth, B. D., y Masters, G. (1982). *Rating scale analysis.* Chicago: MESA.

Anexos

Anexos

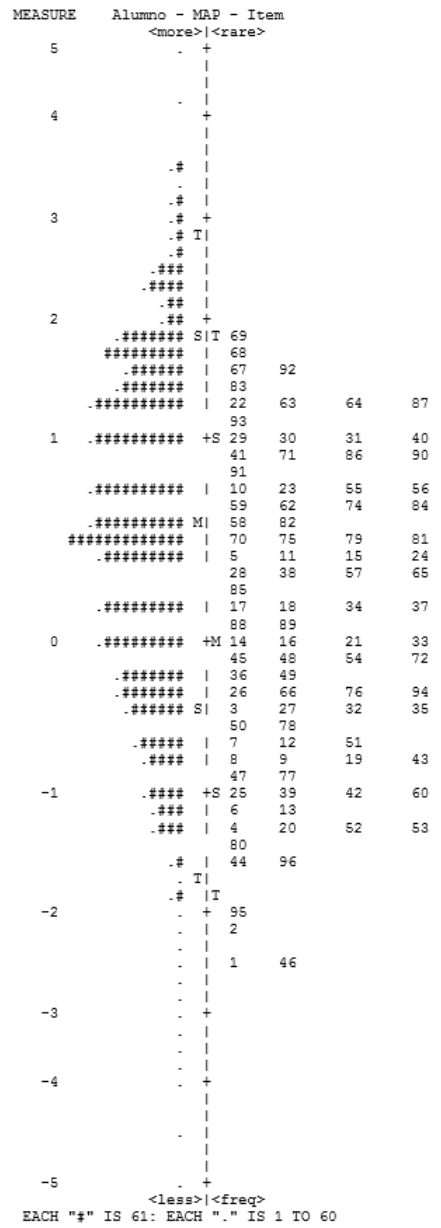

Figura 1. ***Mapa de ítems de Lectura. EM 2018 2.° grado de primaria***

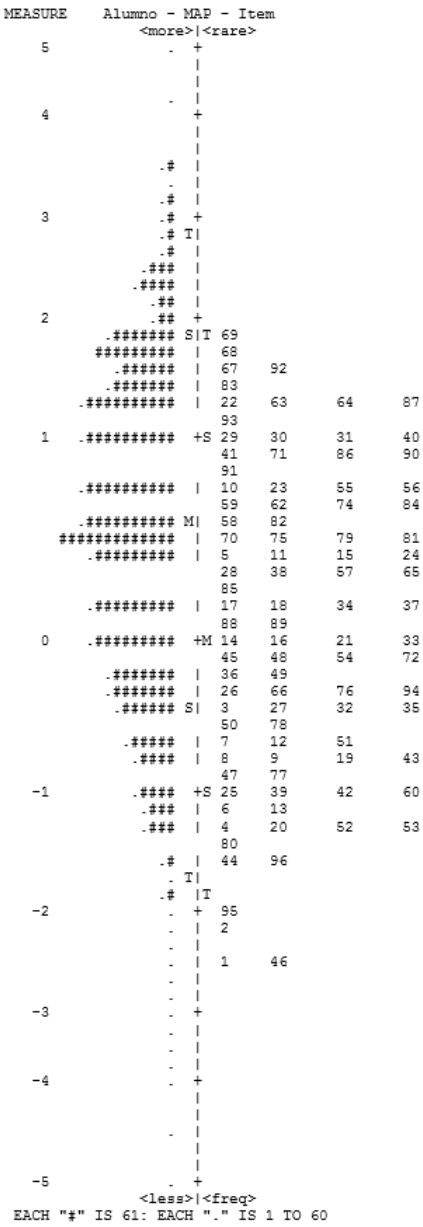

Figura 2. ***Mapa de ítems de Matemática. EM 2018 2.° grado de primaria***

Ministerio de Educación

Calle del Comercio 193,
San Borja - Lima, Perú
Telf: (511) 615-5800

http://www.minedu.gob.pe/

PERÚ | Ministerio de Educación