Banco Interamericano de Desarrollo

Validación de instrumentos de medición utilizados en el proyecto “Tutorías remotas” para acelerar aprendizajes

Juan León
Felipe J. Hevia
Samana Vergara-Lope
Tamara Vinacur
Pablo Zoido

Division de Educación

NOTA TÉCNICA Nº
IDB-TN-02612

Marzo 2022

Validación de instrumentos de medición utilizados en el proyecto “Tutorías remotas” para acelerar aprendizajes

Juan León
Felipe J. Hevia
Samana Vergara-Lope
Tamara Vinacur
Pablo Zoido

Marzo 2022

BID
Banco Interamericano de Desarrollo

Catalogación en la fuente proporcionada por la
Biblioteca Felipe Herrera del
Banco Interamericano de Desarrollo
Validación del instrumento de medición utilizado en el proyecto "Tutorías remotas" para acelerar aprendizajes / Juan León, Felipe J. Hevia, Samana Vergara-Lope, Tamara Vinacur, Pablo Zoido.
p. cm. — (Nota Técnica del BID ; 2612)
1. Educational innovations-Latin America. 2. Educational equalization-Latin America. 3. Mathematics-Study and teaching-Latin America. 4. Educational tests and measurements-Latin America. I. León, Juan. II. Hevia, Felipe. III. Vergara-Lope, Samana. IV. Vinacur, Tamara. V. Zoido, Pablo. VI. Banco Interamericano de Desarrollo. División de Educación. VII. Serie.
IDB-TN-2612

Códigos JEL: I22, I24, I25, I28

Palabras clave: aprendizajes, instrumento de medicion, intervenciones, tutorias remotas

http://www.iadb.org

ÍNDICE

RESUMEN EJECUTIVO

Garantizar resultados de calidad de forma equitativa es el reto principal para los sistemas educativos de América Latina y el Caribe. El principal resultado educativo son los aprendizajes que los estudiantes son capaces de demostrar de forma consistente. No existe ningún instrumento en la región que permita evaluar de forma rápida, costo-efectiva y confiable si una intervención está dando los resultados deseados o no. Con el cierre de escuelas en 2020 y 2021 la necesidad de contar con un instrumento de estas características se volvió aún más urgente. En el contexto de un programa de tutorías remotas para acelerar aprendizajes se adaptaron dos instrumentos preexistentes y se pilotearon en una población de conveniencia para varias jurisdicciones. El objetivo era desarrollar un instrumento económico, de fácil y rápida aplicación que al menos proporcionase una señal confiable de si las intervenciones estaban teniendo un impacto en los aprendizajes fundamentales e indispensables en matemáticas entre la población más vulnerable con mayores brechas de aprendizaje.

El documento analiza dos instrumentos utilizados para evaluar las habilidades matemáticas de los niños y las niñas de 9 a 14 años en números y operaciones, en el marco del Proyecto de Tutorías Remotas para Acelerar Aprendizajes. Estos instrumentos denominados “SMS” y “MIA+” fueron utilizados en varios países de América Latina, con la intención de proporcionar evidencias para estimar el efecto del Programa, y proveer de información pedagógicamente valiosa que contribuya a la mejora en los aprendizajes de los estudiantes que participan en el proyecto.

El instrumento puede ser administrado en forma remota o presencial de forma individual (telefónicamente, por Internet o por un entrevistador) en un tiempo acotado (no más de 20 minutos), por lo que se requiere que con poca cantidad de ítems sea posible contar con evidencias suficientes sobre el nivel de logro por parte de los estudiantes en un conjunto de contenidos de matemática, en función de la cual se personaliza la intervención. Se trata de contar con evidencias que permitan identificar el nivel adecuado de aprendizaje de cada estudiante para que la tutoría pueda constituirse en una oportunidad para remediar y acelerar los aprendizajes.

El análisis realizado proporciona evidencias de que ambos instrumentos (MIA+ y SMS) cuentan con adecuados niveles de confiabilidad (≥0.70) para las muestras de estudio analizadas correspondientes a dos de las jurisdicciones. También se analizó ambas escalas cuentan con una dimensión latente, y una correlación entre el puntaje de la escala SMS y la escala MIA+, que es positiva y significativa, con valores por encima de 0.50, posibilitando la identificación de habilidad matemática en números y operaciones de los estudiantes a partir de la utilización de en ambos instrumentos.

Respecto de la utilidad de la información que proporcionan los instrumentos mencionados, ambos permiten contar con evidencias sobre el efecto de las tutorías remotas. En el caso del

SMS, al tratarse del mismo instrumento utilizado en otros casos, posibilita a su vez la comparación de los resultados obtenidos en América Latina con otros países que han utilizado el modelo de tutorías remotas (como Botsuana, India o Nepal). En el caso de MIA+, también resulta posible indicar el nivel de avance de los estudiantes en distintos ejes de contenido, de modo de propiciar intervenciones oportunas y de forma comparable a otras jurisdicciones donde se aplicaron estos instrumentos anteriormente (principalmente en México). Finalmente, se introducen algunos aspectos a considerar en la mejora del instrumento que pueda aprovechar las mejores cualidades de los instrumentos aquí analizados, aprovechando su potencial para proporcionar información desagregada por eje de contenido y propiciar intervenciones oportunas.

INTRODUCCIÓN

El Proyecto "Tutorías Remotas para Acelerar Aprendizajes" se implementa a partir de 2021 en varios países de América Latina como respuesta ante la presumible pérdida de aprendizajes que vivieron los sistemas educativos en el contexto de COVID 19 con el cierre de las escuelas. Esta iniciativa tiene como objetivo fundamental recuperar los aprendizajes básicos en matemáticas en estudiantes entre 9 y 14 años. Para ello, contempla el desarrollo de tutorías remotas personalizadas y semanales en contenidos de matemática, tomando como referencia el modelo utilizado por Youth Impact (Angrist, Bergman, Brewster y Matsheng, 2020).

El Proyecto adopta el enfoque de Enseñar en el Nivel Adecuado/Teaching at the Right Level (TaRL), desarrollado por Pratham, que sostiene que es necesario trabajar con los estudiantes en función de sus necesidades de aprendizaje (y no necesariamente en función de la edad) para garantizar el aprendizaje de habilidades básicas, a partir de evaluaciones formativas realizadas de manera sistemática que posibiliten el diseño de intervenciones oportunas para el trabajo con cada estudiante en forma personalizada (Banerjee, 2016).

Por este motivo, resulta esencial la identificación del punto de partida en el que se encuentra cada estudiante para poder ofrecer un trayecto formativo acorde a las necesidades singulares de cada alumno. Es por ello por lo que al inicio de la intervención se realiza una evaluación diagnóstica, que se replica al finalizar el proceso de tutorías, permitiendo informar el grado de avance de cada estudiante.

Entre las lecciones aprendidas para evaluaciones de aprendizajes realizadas mediante el teléfono, Angrist et al. (2020) identifican como necesario evaluar la confiabilidad y validez de los instrumentos utilizados. Se sugiere realizar análisis de confiabilidad a través del coeficiente Alpha de Cronbach, análisis estructural o de ítems, utilizando la Teoría de Respuesta al Ítem y análisis de validez concurrente comparando con evaluaciones realizadas en forma presencial.

La evaluación de los aprendizajes en este Proyecto tiene dos objetivos: el primero es identificar el efecto de las tutorías remotas en la mejora de aprendizajes fundamentales de matemática. El segundo, es identificar el nivel correcto o el nivel adecuado de aprendizaje de cada niño para que la tutoría sea efectiva.

1. Los instrumentos de evaluación de aprendizaje

En el contexto de recuperación de los aprendizajes luego de los prolongados cierres de las escuelas producto de la pandemia por COVID-19, existe un fuerte consenso en generar evaluaciones diagnósticas rápidas y sencillas, que procuren entregar información a los docentes, las familias y las autoridades educativas para poder recuperar los aprendizajes perdidos.

En este contexto, surge el proyecto de tutorías remotas para recuperar y acelerar los aprendizajes. El proyecto contempla un primer momento de trabajo con cada sistema educativo realizando la adaptación de la propuesta al contexto local. Se analizaron los marcos curriculares vigentes y se desarrollaron materiales de trabajo, como así también la necesaria alineación con la propuesta de evaluación.

La propuesta de evaluación debe contemplar condiciones similares al modo en que se realiza la propuesta de enseñanza, en el marco de las tutorías. Esto es, se trata de tutorías que se realizan en forma telefónica posibilitando que una mayor cantidad de estudiantes pueda participar en esta iniciativa. Al tratarse de una propuesta de baja complejidad tecnológica, se facilita el acceso de quienes cuentan con dificultades de conectividad, o acceso a equipamiento, e incluso a quienes residen en áreas rurales. Todos los estudiantes pueden participar de las tutorías: solo se requiere contar con un teléfono.

En los distintos encuentros de tutoría abordan contenidos de matemática correspondientes a *Números y operaciones*, que comprende contenidos relacionados con:

- Sistema de numeración,
- Campo aditivo,
- Campo multiplicativo.

En cada uno de estos ejes temáticos, se contempla el trabajo tanto en la resolución de diversos tipos de problemas como en el abordaje de distintas técnicas y estrategias de cálculo mental.

Todos los países en que se realiza el proyecto comparten el recorte curricular de números y operaciones como eje sustantivo en la enseñanza de matemática en el primer y segundo ciclo de la educación primaria. Sin embargo, se presentan algunas variaciones entre países vinculadas a la prioridad que cada sistema educativo le atribuye a aquello que se prioriza como objeto de enseñanza, como así también a algunos aspectos didácticos y a la relación entre los

contenidos entre sí. De todos modos, y a los fines de este análisis, las tres jurisdicciones analizadas comparten las mismas orientaciones para el tutor y los mismos instrumentos de evaluación.

En el diseño de la intervención, se optó por tener dos instrumentos para medir aprendizajes matemáticos fundamentales al inicio y al final del proceso de trabajo con los estudiantes.

SMS

El primero, conocido como SMS, fue una adaptación de los materiales que Youth Impact aplicó en Botsuana. Este instrumento consta de nueve reactivos de los cuales se mide valor posicional, suma, multiplicación, división, y problemas lógicos. Para cada reactivo se da solo una respuesta posible y se tiene que responder en menos de 2 minutos cada uno. Para la identificación del nivel adecuado, se utilizan los cinco primeros reactivos (valor posicional, suma, resta, multiplicación, división) (Anexo 1). Se aplican todos los ítems uno a uno y se dan respuestas de logro-error.

MIA +

El segundo instrumento se denomina MIA plus, fue desarrollado por el proyecto MIA para el desarrollo de intervenciones educativas basadas en el principio de enseñar a nivel adecuado en México y es una adaptación (Hevia, Vergara-Lope & Velásquez-Durán, 2022) del instrumento MIA propuesto originalmente por Hevia y Vergara-Lope (2016). Este instrumento mide operaciones matemáticas básicas y resolución de problemas. Consta de nueve reactivos: número, suma sin acarreo, suma con acarreo, resta sin transformación, restas con transformación, división, problema con apoyo visual, problema sin apoyo visual y fracciones. Para el caso de los reactivos de operaciones, los sujetos tienen que responder al menos dos de tres opciones de manera correcta; y para los reactivos de problemas, el sujeto puede responder correctamente hasta en dos oportunidades. Para la identificación del nivel adecuado para las tutorías se utilizan los seis primeros reactivos (número, suma uno, suma dos, resto a uno, resta dos, división) (Anexo 2). Este instrumento aplica los ítems por nivel de dificultad y descontinúa su aplicación al momento en el que el sujeto no puede responder uno de los ítems.

A continuación, se comparan ambos instrumentos:

Cuadro 1. Instrumentos de medición proyecto tutorías remotas para acelerar aprendizajes

Contenido evaluado		Ejemplo de ítem	Nº reactivos	
			SMS	MIA +
Sistema de numeración	Lectura y ordenamiento de números	Puede leer números		1
	Valor posicional	Decodifica números	1	

Campo aditivo	Estrategias de cálculo	Suma sin acarreo		1
		Suma 2 dígitos con acarreo	1	1
		Resta 2 dígitos sin acarreo		1
		Resta 2 dígitos con acarreo	1	1
	Resolución de problemas de suma y resta			1
Campo multiplicativo	Estrategias de cálculo	Resuelve multiplicaciones	1	
		Resuelve divisiones	1	1
	Resolución de problemas	Resuelve problemas		1
		Resuelve fracciones		1
Otros ítems	Pensamiento lógico	Resuelve problemas lógicos	4	
	Total ítems		9	9

Fuente: Elaboración propia

Tal como se desprende del cuadro anterior, en ambos casos se trata de instrumentos de evaluación de corta extensión, lo cual permite ser administrado en forma telefónica, manteniendo la atención de los estudiantes a lo largo del tiempo en que se realiza la prueba. A su vez, la distribución de ítems según eje de contenido permite dar cuenta de la prioridad asignada a cada uno de los mismos, y de las oportunidades de proporcionar evidencias más exhaustivas respecto del grado de avance de los estudiantes en cada uno de los aspectos seleccionados. Se observa que el diseño de MIA + cubre de manera mas exhaustiva los distintos ejes de contenido priorizados, lo cual favorece un mejor identificación de la situación inicial de cada estudiante.

2. Alcances, preguntas y metodología

El presente informe tiene como objetivo principal evaluar la confiabilidad y las evidencias de validez de los módulos del MIA+ y SMS administrados a estudiantes entre 9 y 14 años que pueden estar asistiendo a distintos grados/ años de escolaridad.

Resulta necesario avanzar hacia la construcción de un único instrumento que pueda utilizarse para proporcionar evidencias sobre: a) el efecto de las tutorías telefónicas en la mejora de aprendizajes fundamentales de matemáticas; y b) proporcionar información confiable para identificar los logros y dificultades persistentes que presenta cada estudiante en Números y Operaciones, de modo de propiciar intervenciones oportunas tanto durante la tutoría, como para compartir los avances con la escuela y/o el Ministerio de Educación correspondiente.

Dadas las características del Proyecto, se debe tratar de un instrumento que sea factible de ser administrado en forma remota (por medio del teléfono y eventualmente, a través de internet). Además, debe ser breve, de modo de procurar sostener la atención de los estudiantes a lo largo del tiempo previsto para la evaluación, y debe proporcionar evidencias sobre los distintos ejes de contenido mencionados, que son los aspectos sobre los cuales se centra la propuesta de enseñanza. Algunos de estos desafíos son identificados también por Angrist et al. (2020) como lecciones aprendidas, incorporando a su vez la necesidad de propiciar la construcción de un clima de confianza y compromiso con los estudiantes y sus cuidadores, que posibilite realizar evaluaciones breves adaptadas al formato oral, que provean información sobre los aprendizajes de los estudiantes de un modo costo efectivo.

El análisis realizado utiliza los datos correspondientes a la aplicación de ambos instrumentos en tres jurisdicciones: Jurisdicción A, Jurisdicción B y Jurisdicción C. Se analizó, en primer lugar, la consistencia interna (confiabilidad) del módulo MIA+ y SMS. Posteriormente, se evaluó la evidencia de validez del módulo MIA+ y SMS. De manera adicional, una vez obtenidos los puntajes se procedió a analizar la distribución de estos y correlacionarlo con diferentes variables demográficas para poder apreciar su validez predictiva.

De este modo se intentó responder a las siguientes preguntas:

a) ¿Son confiables los instrumentos (consistencia interna)?
b) ¿En qué medida los instrumentos permiten proporcionar evidencias sobre el efecto de las tutorías remotas en los aprendizajes de los estudiantes? ¿Y para brindar información de los distintos ejes de contenido, de modo de propiciar intervenciones oportunas?
c) ¿Habría que mejorar algunos de los ítems para fortalecer los instrumentos que se están utilizando actualmente?
d) ¿Qué evidencias hay hasta el momento de la validez predictiva de los instrumentos? (en relación a otras escalas con las que habitualmente se asocia el constructo evaluado)

El análisis realizado permitió conocer las fortalezas y oportunidades de mejora de cada instrumento, con la intención de avanzar hacia el diseño de un nuevo instrumento que proporcione respuestas a los desafíos identificados.

2.1 Muestra

Para la elaboración de la muestra, se solicitó a las autoridades educativas de cada jurisdicción un listado de 10,000 estudiantes que cumplieran con los requisitos de estar en las edades señaladas para este estudio (entre 9 y 14 años) y estar inscritos en escuelas públicas. De estos casos, se procedió a hacer una muestra aleatoria de cerca de 3,200 casos por jurisdicción.

El cuadro 1 presenta la distribución de la muestra bajo estudio de acuerdo con el sexo y edad. De esta manera, se puede apreciar que se cuenta con una muestra balanceada por sexo; mientras que por edad, no se cuenta con este mismo aspecto. En el caso de edad, se cuenta

con un mayor porcentaje de estudiantes entre 11 y 14 años para la jurisdicción A, y en el caso de las jurisdicciones B y C, el mayor porcentaje de evaluados está entre los 9 y 11 años.

Cuadro 2. Niños y niñas evaluados por sexo, edad y grado de estudios.

	Jurisdicción A		Jurisdicción B		Jurisdicción C	
	N	%	N	%	N	%
Total	3440	100.0	3210	100.0	3238	100.0
Sexo						
Mujeres	1782	51.8	1594	49.7	1613	49.8
Hombres	1658	48.2	1615	50.3	1625	50.2
Edad						
9	305	8.9	510	15.9	797	24.6
10	554	16.1	756	23.6	718	22.2
11	618	18.0	1151	35.9	794	24.5
12	654	19.0	454	14.1	619	19.1
13	622	18.1	187	5.8	212	6.6
14	687	20.0	152	4.7	98	3.0

Fuente: Elaboración propia a partir de las bases de datos de línea de base 2021.

2.2 Análisis de Confiabilidad

El análisis de confiabilidad permite establecer qué tan bien una prueba o escala mide el constructo que se supone está midiendo; en otras palabras, lo que se busca medir es la consistencia con la que se mide un constructo. Los métodos principales para medir la confiabilidad de una escala o prueba son dos: el test-retest y el análisis de consistencia interna. El test-retest consiste en tomar la escala o prueba al individuo en dos oportunidades y calcular la correlación entre ambos puntajes. En el caso del análisis de consistencia interna, lo que se mide es el grado en el que los ítems que se incluyen dentro de una escala se correlacionan entre sí, es decir, están midiendo el mismo constructo. Por lo general, se hace uso de índices como el Alpha de Cronbach o el de Kuder-Richardson 20 para medir el nivel de consistencia interna de una escala o prueba.

Para los análisis de confiabilidad realizados se hace uso del índice del Alpha de Cronbach cuya fórmula es la siguiente:

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_x^2}\right]$$

α = índice de confiabilidad

k = número de ítems en la escala o prueba

σ_i^2 = varianza del ítem i

σ_x^2 = varianza de la escala o prueba

Este índice toma valores de 0 a 1. Valores cercanos a 0 indican que hay poca confiabilidad o consistencia interna entre los ítems de la escala o prueba, mientras valores cercanos a 1 indican que los ítems tienen una alta consistencia interna y estarían midiendo el mismo constructo. De acuerdo a la literatura, como mencionamos anteriormente, valores iguales o mayores a 0.70 indican que los ítems de la escala estarían midiendo de manera adecuada el mismo constructo (Nunnally & Bernstein, 1994).

Sin embargo, si bien el alfa de Cronbach es el índice de confiabilidad más utilizado, existen críticas con respecto a su uso para medir la consistencia interna de una prueba o escala. Por este motivo, se estima también el índice de confiabilidad Omega (McDonald, 1999) dado que cuenta con menores sesgos que los planteados al alfa de Cronbach (Cho & Kim, 2015). La fórmula para su cálculo es:

$$\omega = \frac{(\sum_{i=1}^{k}\lambda_i)^2}{(\sum_{i=1}^{k}\lambda_i)^2 + \sum_{i=1}^{k}1 - \lambda_i^2}$$

Donde λ_i es el peso de cada ítem en un modelo de análisis factorial de un solo factor. Mientras $1 - \lambda_i$ representa a la varianza única de cada ítem. Finalmente, al igual que el alfa de Cronbach, el índice Omega oscila entre 0 y 1, donde valores cercanos a 0 indican ausencia de confiabilidad de la escala analizada, y valores cercanos a 1 indican una alta confiabilidad de la escala analizada.

2.3 Evidencia de validez

El análisis de validez consiste en verificar si una escala o prueba mide las dimensiones que se supone está midiendo. Los tipos de evidencia de validez comúnmente usados son:

i) *Validez de contenido:* método que permite validar una escala o prueba por medio de expertos o especialistas en el área, quienes evalúan la adecuación e idoneidad de la escala o prueba para medir el constructo propuesto.

ii) *Validez concurrente o predictiva:* método que permite validar una escala o prueba mediante su correlación con otra escala o prueba que mida el mismo constructo, asimismo mediante la correlación de la escala con otras variables que de acuerdo a la literatura están asociadas con el constructo que se busca medir.

iii) *Validez de constructo:* método mediante el cual se verifica la relación entre los ítems o preguntas que conforman la escala o prueba con el constructo teórico que se supone está midiendo.

Para el presente informe, se va verificó la *validez de constructo* para el módulo MIA+ (muestra jurisdicción B) y SMS (muestra jurisdicciones A, B y C). En el caso del MIA+ se exploró el número de dimensiones que hay en los ítems que conforman el módulo para validarlo; mientras que en el SMS se validó la existencia de un solo constructo matemático. Para realizar este ejercicio, se hizo uso del Análisis Factorial Exploratorio -AFE- y el Análisis Factorial Confirmatorio -AFC- (Nunnally & Bernstein, 1994; Howell, 2006). El primero permite explorar el número de factores latentes detrás de un conjunto de variables y el segundo sirve para validar la estructura factorial que se plantea teóricamente en un instrumento desarrollado. Por otro lado, dada la métrica de los ítems (dicotómicos) en la escala del MIA+ y SMS, se estimó el AFE y AFC usando la matriz de correlaciones tetracórica que es la más adecuada para ítems de tipo dicotómico o binarios.

Finalmente, para ver que se cuenta con un buen ajuste en el modelo EFA, se estimó el indicador de *Kaiser-Meyer-Oklim* que permite medir la adecuación muestral de los datos al modelo planteado y valores por encima de 0.70 indican buen ajuste (Kaiser. 1970). En el caso de los modelos AFC, se emplearon dos medidas de ajuste comparativo y dos medidas de ajuste absoluto. Estas medidas son:

i) CFI (índice de ajuste comparativo): compara el ajuste de modelos en base a un modelo base o de independencia. Cuanto más se acerque a 1, se considera que el modelo ajusta

bien. Se considera un buen ajuste cuando el valor es igual o mayor a 0.90 (Hair et al., 2006).

ii) TLI (índice de Tucker Lewis): a diferencia del CFI permite penalizar por la inclusión de parámetros en el modelo; al igual que el CFI, valores cercanos a 1 indican un buen ajuste. Se considera un buen ajuste cuando el valor es igual o mayor a 0.90 (Hair et al., 2006).

iii) SRMR (la raíz cuadrática media de los residuos estandarizados): viene a ser la normalización de la diferencia entre la correlación observada y la correlación pronosticada por el modelo. Un valor de cero indica ajuste perfecto y un valor inferior a 0.08 se considera un buen ajuste (Hu y Bentler, 1999).

iv) RMSEA (error cuadrático medio de aproximación): indica qué tan bien los parámetros de un modelo reproducirán las covarianzas poblacionales. Un modelo que estime exactamente estas covarianzas tendrá un RMSEA de 0, por lo que se requiere que el valor de este estadístico sea bajo para poder decir que el modelo tiene un buen ajuste. Se considera un buen ajuste cuando el valor es igual o menor a 0.06 (Hair et al., 2006).

Con los resultados obtenidos se procedió a realizar un segundo procedimiento orientado a la validez concurrente, utilizando para ello análisis de correlaciones y regresiones lineales simples, además de identificar la correlación del puntaje estimado para la escala SMS y variables demográficas como el sexo, edad y nivel socioeconómico.

2.4 Teoría de Respuesta al Ítem (TRI)

En la TRI, se asume que la habilidad de un individuo depende de tres parámetros: i) dificultad del ítem (b_i), ii) discriminación del ítem (a_i), y iii) el azar (c_i). Estos tres parámetros son relacionados mediante una función de distribución logística como se muestra a continuación:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}} \qquad i = 1,2,\dots,n$$

$P_i(\theta)$: la probabilidad que un individuo con habilidad θ acierte en el ítem i

a_i : ítem de discriminación

b_i : ítem de dificultad

c_i : probabilidad de acertar el ítem al azar

n : el número de ítems en el test

θ : el parámetro de habilidad del individuo

La ecuación anterior muestra la función general de lo que sería el modelo de tres parámetros (3PL). Sin embargo, si asumimos que no existe el azar o adivinación para acertar un ítem en la prueba ($c_i=0$), la probabilidad de acertar correctamente dependería sólo de la dificultad y la discriminación del ítem, que sería el modelo de 2 parámetros (2PL). Finalmente, si asumimos, que la discriminación de los ítems es igual para todos ($a_i=1$), se tiene que la probabilidad de responder correctamente a un ítem dependería solo de su dificultad y tendríamos el modelo de un parámetro (1PL) o Modelo Rasch (Crocker & Algina, 1986).

Las ventajas que tiene usar este último modelo son: i) permite que los estadísticos de ajuste de los ítems no dependan de la muestra que está siendo evaluada, ii) permite establecer una relación clara entre la dificultad de un ítem y el puntaje de los individuos, iii) dado que se basa en el ítem facilita la equiparación de puntajes entre pruebas, y iv) los valores perdidos son manejados fácilmente dado que se basan en toda la información disponible en un ítem y no de la prueba.

3. Resultados

En el presente acápite se muestran los resultados de los análisis realizados. En primer lugar, se presentan los análisis de confiabilidad para cada uno de los módulos. En segundo lugar, se presenta la evidencia de validez de constructo para el módulo SMS. Luego, se presentan los análisis realizados usando TRI a la escala SMS. Por último, se describen los procedimientos utilizados para identificar el nivel adecuado de aprendizaje.

3.1. Confiabilidad

Los resultados muestran que ambos instrumentos (SMS y MIA+) poseen una adecuada discriminación interna e índices de confiabilidad adecuados.

3.1.1. Dificultad y discriminación de los ítems (Teoría Clásica)

Un primer aspecto que se exploró es la tasa de acierto o nivel de dificultad de cada uno de los ítems para ambas escalas, así como el nivel de discriminación de cada ítem. En el caso de este último indicador, se trata de qué tan bien cada ítem permite discriminar entre aquellos estudiantes que tienen una mayor o menor habilidad en el constructo que se está midiendo. En el caso del MIA+, se aprecia en el siguiente cuadro que existe una variación grande en la tasa de acierto o dificultad de los ítems para la muestra de la jurisdicción A. Se puede observar que los cuatro últimos ítems cuentan con tasas de acierto menores al 10% e, incluso, los últimos tres ítems tienen tasas de acierto menores al 5%. De la misma forma, se aprecia que en el caso de uno de los ítems, más del 90% de los niños y niñas evaluados lo respondieron correctamente. En el caso de las jurisdicciones B y C, se puede apreciar una menor variación en las tasas de acierto, pero de igual forma se observa que son los primeros ítems los más

fáciles y los últimos los más difíciles; pero a diferencia de la jurisdicción A, los ítems más difíciles son respondidos por un poco más del 10% de los y las estudiantes evaluados. Si bien lo que se busca es tener variación en la tasa de acierto o dificultad de los ítems incluidos en una evaluación, no se debe tener ítems con una variabilidad muy baja dado que no van a poder brindar mucha información sobre la habilidad de los y las estudiantes.

En cuanto a los niveles de discriminación de los ítems, se aprecia que estos son adecuados en todos los casos para las tres muestras de estudio, contando con correlaciones ítem-resto de la prueba mayores a 0.20. Cuadro 3. Índices de dificultad y discriminación de los ítems usados en el módulo MIA+

	Jurisdicción A (n=3440)		Jurisdicción B (n=3210)		Jurisdicción C (n=3238)	
	Dificultad	Discriminación[1/]	Dificultad	Discriminación[1/]	Dificultad	Discriminación[1/]
math_plus1	0.916	0.351	0.955	0.254	0.935	0.371
math_plus2	0.777	0.516	0.914	0.374	0.859	0.517
math_plus3	0.556	0.679	0.802	0.534	0.733	0.664
math_plus4	0.399	0.728	0.755	0.563	0.639	0.725
math_plus5	0.281	0.682	0.642	0.602	0.541	0.734
math_plus6	0.088	0.527	0.409	0.590	0.287	0.718
math_plus7	0.047	0.482	0.374	0.595	0.216	0.704
math_plus8	0.033	0.446	0.242	0.514	0.170	0.661
math_plus9	0.025	0.398	0.308	0.523	0.141	0.605

Fuente: Elaboración propia a partir de las bases de datos de línea de base 2021.

1/ Se usó la correlación ítem-resto de la prueba

En el caso del módulo SMS, a diferencia del MIA+, se aprecia que los ítems no han sido medianamente difíciles en promedio para los y las estudiantes evaluados en las diferentes muestras. Se aprecia que la dificultad de los ítems no es muy grande, oscilando entre 0.10 (math_sms5) y 0.46 (math_sms2) para el caso de la jurisdicción A, y además no hay ningún ítem que haya sido respondido por el 50% o más de los y las estudiantes. En el caso de la jurisdicción C, la tasa de acierto oscila entre 0.26 (math_sms9) y 0.76 (math_sms2), y sí se cuenta con ítems que han sido respondidos por más del 50% de los y las evaluados. Para la jurisdicción B, se puede apreciar que la tasa de acierto osciló entre 0.18 (math_sms9) y 0.71 (math_sms2), y se cuenta con ítems que, incluso, los respondieron correctamente más de las dos terceras partes de los y las estudiantes evaluados. Finalmente, en el caso de la

discriminación de los ítems, se aprecia que esta es adecuada para todas las muestras, teniéndose que las correlaciones ítem-resto de la prueba están por encima de 0.20 en todos los casos.

Cuadro 4. Dificultad y discriminación de los ítems usados en el módulo SMS

	Jurisdicción A (n=3440)		Jurisdicción B (n=3210)		Jurisdicción C (n=3238)	
	Dificultad	Discriminación[1/]	Dificultad	Discriminación[1/]	Dificultad	Discriminación[1/]
math_sms1	0.311	0.426	0.403	0.413	0.637	0.494
math_sms2	0.460	0.507	0.706	0.487	0.760	0.445
math_sms3	0.273	0.565	0.466	0.535	0.586	0.543
math_sms4	0.208	0.560	0.504	0.585	0.522	0.563
math_sms5	0.104	0.411	0.236	0.487	0.326	0.523
math_sms6	0.450	0.475	0.635	0.509	0.755	0.465
math_sms7	0.363	0.223	0.233	0.282	0.377	0.354
math_sms9[2/]	0.170	0.340	0.175	0.257	0.260	0.376
math_sms10	0.299	0.441	0.498	0.404	0.522	0.423

Fuente: Elaboración propia a partir de las bases de datos de línea de base 2021.

1/ Se uso la correlación ítem-resto de la prueba

2/ Se cuenta con 1285 estudiantes en la jurisdicción A, 1169 en la jurisdicción B y 1679 en la jurisdicción C.

3.1.2. Índices de Confiabilidad

Una vez revisada la dificultad y discriminación de cada uno de los ítems, se procedió a estimar la confiabilidad. Tal como se mencionó en la sección previa, se hace uso de dos índices comúnmente usados para ver la confiabilidad de escalas o pruebas como es el índice de *Cronbach* y el índice de *Omega* (McDonald, 2013). En el siguiente cuadro, se aprecia que, para ambos módulos, los niveles de confiablidad, tanto de *Cronbach* u *Omega*, son adecuados y están por encima de 0.70 para todas las muestras, apreciándose de esta manera que se cuenta con pruebas que brindan confianza en la replicabilidad de sus resultados.

De manera adicional, se hizo el mismo ejercicio, pero solo para un grupo reducido de ítems en ambos módulos, en el caso del MIA+ para los cinco primeros y en el SMS para los seis primeros. La idea detrás de hacer este sub-análisis responde al hecho de que estamos considerando solo las habilidades básicas que deberían de tener los y las estudiantes en materia de la competencia numérica. Los resultados muestran (ver cuadro 5) que, al igual que la escala completa, para todas las muestras se cuentan con niveles de confiabilidad adecuados para medir las habilidades matemáticas básicas de los y las estudiantes.

Cuadro 5. Confiabilidad de los módulos administrados.

	Jurisdicción A (n=3440)		Jurisdicción B (n=3210)		Jurisdicción C (n=3238)	
	Cronbach	Omega	Cronbach	Omega	Cronbach	Omega
MIA+						
Todos los ítems	0.817	0.820	0.813	0.810	0.885	0.851
Seis primeros ítems	0.819	0.812	0.748	0.757	0.851	0.861
SMS						
Todos los ítems	0.766	0.778	0.764	0.766	0.775	0.777
Cinco primeros ítems	0.741	0.751	0.720	0.724	0.734	0.735

Fuente: Elaboración propia a partir de las bases de datos de línea de base 2021.

3.2. Validez

Respecto a la validez de constructo, el módulo SMS muestra unidimensionalidad. En el caso de MIA+ se encuentra una estructura de tres factores con indicadores de ajuste adecuados, resultado esperable de acuerdo con sus objetivos: indicar el nivel de avance que tienen los y las estudiantes en las habilidades numéricas.

3.2.1. Validez de constructo

Antes de revisar los análisis de validez de constructo, un primer aspecto que hay que tomar en consideración es el diseño de cada evaluación o módulo. En el caso del módulo SMS, los ítems buscan evaluar la competencia numérica que tienen los y las estudiantes de manera global; sin embargo, a diferencia del módulo MIA+ administrado en las jurisdicciones B y C, los y las estudiantes respondieron a todos los ítems con un tiempo determinado, pero sin condicionar la respuesta de un ítem a otro del mismo módulo (salvo en las preguntas 8 y 9). Este aspecto hace que los ítems en cierta forma sean independientes uno de otro y se pueda estimar tranquilamente el constructo latente detrás del conjunto de ítems evaluados.

Se hizo el ejercicio de estimar el análisis factorial confirmatorio probando que existe un solo factor latente detrás de todos los ítems. Por otro lado, se tomó en consideración la métrica de los ítems, por lo que el análisis factorial se realiza usando la matriz de correlaciones tetracórica entre los ítems. No se consideraron los ítems 8 y 9 porque el ítem 8 es una pregunta de control y el ítem 9 reduce la muestra de forma considerable en cada una de las muestras bajo estudio, motivo por el cual de los diez ítems que tiene la escala solo se usaron en los análisis ocho de ellos. Los resultados que se aprecian en el siguiente cuadro indican que tanto a nivel individual (ítem) como a nivel del modelo se cuenta con un buen ajuste y se sostiene la existencia de un solo factor latente en las tres muestras bajo estudio[1].

[1] En el Anexo 1 se cuenta con los cuadros del AFC usando la métrica continua y se puede apreciar que no existen mayores diferencias en los resultados obtenidos.

Cuadro 6. Análisis factorial confirmatorio del módulo SMS, escala completa

Jurisdicción A			
	Peso factorial	P-value	R2
math_sms1	0.633	0.000	0.401
math_sms2	0.807	0.000	0.652
math_sms3	0.840	0.000	0.706
math_sms4	0.869	0.000	0.755
math_sms5	0.789	0.000	0.623
math_sms6	0.770	0.000	0.592
math_sms7	0.352	0.000	0.124
math_sms10	0.576	0.000	0.332

RMSEA	0.130
CFI	0.924
TLI	0.893
SRMR	0.040

Jurisdicción B			
	Peso factorial	P-value	R2
math_sms1	0.574	0.000	0.330
math_sms2	0.747	0.000	0.558
math_sms3	0.778	0.000	0.606
math_sms4	0.846	0.000	0.717
math_sms5	0.773	0.000	0.598
math_sms6	0.751	0.000	0.564
math_sms7	0.422	0.000	0.178
math_sms10	0.559	0.000	0.312

RMSEA	0.105
CFI	0.939
TLI	0.915
SRMR	0.042

Jurisdicción C			
	Peso factorial	P-value	R2
math_sms1	0.708	0.000	0.501
math_sms2	0.702	0.000	0.493
math_sms3	0.795	0.000	0.632
math_sms4	0.796	0.000	0.634
math_sms5	0.770	0.000	0.593
math_sms6	0.713	0.000	0.509
math_sms7	0.496	0.000	0.246
math_sms10	0.580	0.000	0.336

RMSEA	0.155
CFI	0.877
TLI	0.827
SRMR	0.066

Fuente: Elaboración propia a partir de las bases de datos de línea de base 2021.

Finalmente, se hizo el mismo ejercicio para la versión reducida del SMS. Al igual que lo presentado para la escala completa, se aprecia que se sostiene el modelo de un solo factor para la escala reducida dado que los indicadores de ajuste global (CFI, TLI, RMSEA y SRMR) del modelo planteado cumplen con los criterios para decir que se cuenta con un buen ajuste en el modelo.

Cuadro 7. Análisis factorial confirmatorio del módulo SMS, escala reducida

Jurisdicción A			
	Peso factorial	P-value	R2
math_sms1	0.637	0.000	0.405
math_sms2	0.794	0.000	0.630
math_sms3	0.838	0.000	0.702
math_sms4	0.885	0.000	0.784
math_sms5	0.789	0.000	0.622

RMSEA	0.211
CFI	0.927
TLI	0.854
SRMR	0.042

Jurisdicción B			
	Peso factorial	P-value	R2
math_sms1	0.564	0.000	0.319
math_sms2	0.728	0.000	0.530
math_sms3	0.781	0.000	0.609
math_sms4	0.866	0.000	0.751
math_sms5	0.770	0.000	0.594

RMSEA	0.131
CFI	0.962
TLI	0.925
SRMR	0.032

Jurisdicción C			
	Peso factorial	P-value	R2
math_sms1	0.702	0.000	0.493
math_sms2	0.694	0.000	0.481
math_sms3	0.819	0.000	0.671
math_sms4	0.787	0.000	0.619
math_sms5	0.764	0.000	0.584

RMSEA	0.142
CFI	0.956
TLI	0.912
SRMR	0.035

Fuente: Elaboración propia a partir de las bases de datos de línea de base 2021.

En el caso del módulo MIA+, este tiene la finalidad de poder indicar el nivel de avance que tienen los y las estudiantes en las habilidades numéricas; en otras palabras, si son capaces solo de identificar números o de resolver ejercicios de fracciones. Dado esto, la evaluación realizada desde su diseño hace que exista dependencia entre los ítems administrados y por ende el puntaje que se obtenga no será un rasgo latente sino una escala ordinal donde el valor indica hasta qué aspecto numérico (por ej.: sumas) puede llegar a resolver el o la estudiante. Por lo tanto, en el caso de la escala MIA+ no tendría sentido realizar el análisis de validez de constructo dado que por diseño se condicionó la respuesta de un ítem y otro para las muestras de la jurisdicción A y de la jurisdicción C. Sin embargo, para efectos de evaluación de los instrumentos de medición en el caso de la jurisdicción B, se administró el instrumento sin esta regla y cada estudiante evaluado respondió a todos los ítems de la escala. Este hecho permite revisar si existen evidencias de validez en la escala MIA+ para esta muestra y verificar que se formen efectivamente los constructos o dimensiones esperadas.

Para realizar el ejercicio con la muestra de la jurisdicción B para el MIA+, primero se dividió de forma aleatoria a la muestra de la jurisdicción B en dos mitades. En la primera mitad se realizó

el Análisis Factorial Exploratorio para ver cuántos factores se forman, usándose para la extracción de los factores el método de ejes principales y para la rotación de estos una oblicua que considera correlación entre los factores. El siguiente cuadro muestra que se han formado cuatro factores; sin embargo, los tres primeros reflejan el ordenamiento de los ítems esperado. Es decir, se tiene que la decodificación de números y la suma sin llevar se agrupan en un solo factor, luego en un segundo factor tenemos la suma llevando y las restas con y sin llevar; y finalmente, en un tercer factor tenemos las divisiones, resolución de problemas y fracciones. En cierta forma podemos ver cómo se agrupan los ítems de acuerdo con lo esperado y se aprecia que los tres primeros factores cuentan con una correlación mediana ($r \geq 0.50$), mientras la correlación con el último factor es baja.

Cuadro 8 Resultados del Análisis factorial confirmatorio con rotación oblicua y la correlación entre los factores.

	Varianza			
	4.60	**4.55**	**3.69**	**0.33**
math_plus1			0.798	
math_plus2			0.699	0.204
math_plus3	0.239	0.372	0.267	0.283
math_plus4		0.802		
math_plus5		0.747		
math_plus6	0.637	0.290		
math_plus7	0.798			
math_plus8	0.890			
math_plus9	0.668		0.269	

Matriz de correlaciones				
	Factor 1	**Factor 2**	**Factor 3**	**Factor 4**
Factor 1	1			
Factor 2	0.6797	1		
Factor 3	0.5325	0.6418	1	
Factor 4	0.2946	0.3519	0.2956	1

Nota: Solo se incluyen las cargas factoriales superiores a |0.20| en la matriz de componentes.

Fuente: Elaboración propia a partir de las bases de datos de línea de base 2021.

Luego de obtenida la posible estructura factorial, se procedió a realizar con la segunda muestra el Análisis Factorial Confirmatorio. Se evaluó que en la escala se formaran los tres factores encontrados. Para lo cual se procedió a usar la matriz de correlación tetracórica entre los ítems que son parte de la escala MIA+. Los resultados se aprecian en el siguiente cuadro. Se puede ver que la estructura factorial de tres factores es adecuada y cuenta con indicadores de ajuste

tanto absolutos (RMSEA y SRMR) como relativos bastante adecuados (CFI y TLI); con lo cual se confirma que la estructura factorial encontrada se ajusta a los datos obtenidos.

Cuadro 9. Análisis factorial confirmatorio del módulo MIA+, escala completa

	Peso factorial	P-value	R2
Factor 1			
math_plus1	0.500	0.000	0.250
math_plus2	0.817	0.000	0.668
Factor 2			
math_plus3	0.672	0.000	0.452
math_plus4	0.729	0.000	0.531
math_plus5	0.717	0.000	0.514
Factor 3			
math_plus6	0.692	0.000	0.479
math_plus7	0.763	0.000	0.583
math_plus8	0.674	0.000	0.454
math_plus9	0.639	0.000	0.409

RMSEA	0.062
CFI	0.964
TLI	0.946

Fuente: Elaboración propia a partir de las bases de datos de línea de base 2021.

Finalmente, a manera de validar que se cuenta con una adecuada estructura factorial, se estimó el AFC asumiendo que existe una sola dimensión en los ítems del módulo MIA+ como modelo competitivo al que se encontró en el análisis factorial. Como se aprecia en el siguiente cuadro, los resultados de los indicadores de ajuste validan que la mejor estructura para el módulo es de tres dimensiones.

Cuadro 10 Indicadores de ajuste del análisis AFC asumiendo una y tres dimensiones

	RMSEA	SRMR	CFI	TLI
3 factores	0.062	0.038	0.964	0.946
1 factor	0.142	0.086	0.786	0.715
Variación	-0.08	-0.048	0.178	0.231

Fuente: Elaboración propia a partir de las bases de datos de línea de base 2021.

A manera de resumen de este apartado, en el siguiente cuadro se sintetiza lo encontrado en el análisis de validez de constructo para el módulo SMS. Se aprecia que tanto para la versión completa como para la versión reducida, se sostiene el modelo de un solo factor latente. Por

otro lado, en el caso del análisis de validez realizado a la muestra de la jurisdicción B, se pudo apreciar que el módulo cuenta con tres factores latentes que diferencia ítems de acuerdo con los niveles de dificultad de estos. Este último resultado permite validar la forma de calificación usada para el módulo MIA+ donde, para la muestra de las jurisdicciones A y C, se cuenta con una escala ordinal donde su puntaje nos indica hasta dónde pueden llegar los estudiantes. Finalmente, se pueden emplear los ítems del módulo SMS para complementar la secuencia de dificultad en los ítems y tener una mayor graduación en las habilidades matemáticas que se miden en los y las estudiantes.

Cuadro 11. Resultados del análisis de validez de constructo

	Completa	Reducida
MIA+ (Jurisdicción B)	No sostiene 1 factor Sostiene 3 factores	-
SMS	Sostiene 1 factor	Sostiene 1 factor

3.2.2. Niveles de habilidad en el MIA+

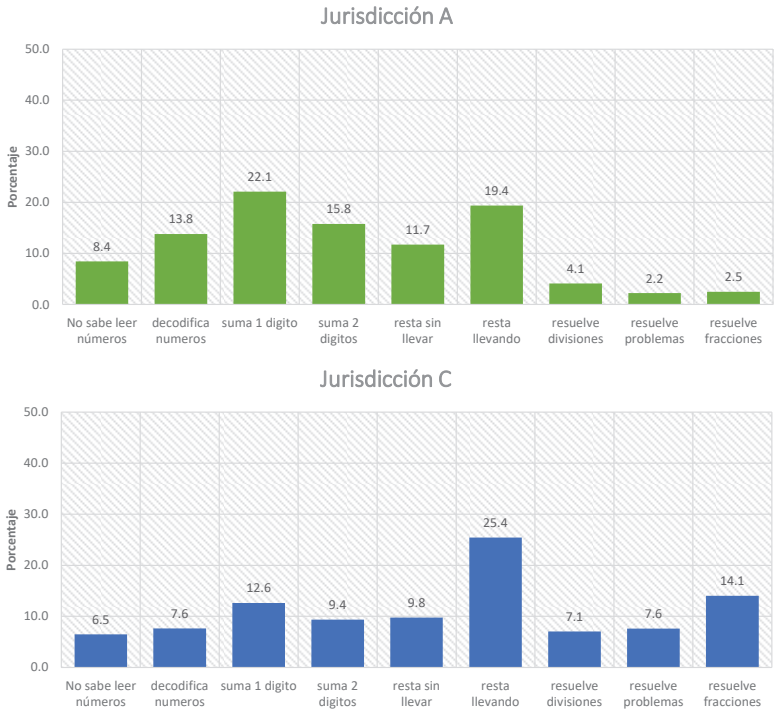
Para complementar el análisis de constructo en el caso de MIA+ con los resultados obtenidos en las jurisdicciones A y C, se procedió a realizar un análisis descriptivo en función de los puntajes obtenidos por los y las estudiantes de estas jurisdicciones. De esta manera, como se mencionó en la sección previa, los puntajes de esta escala ya indican hasta qué capacidad numérica llegan responder los y las estudiantes. Así, en el siguiente cuadro se puede apreciar la distribución de los evaluados de acuerdo al puntaje obtenido en el MIA+. Como se puede observar, existe una gran diferencia entre la muestra de línea base de las jurisdicciones A y C, siendo en esta última donde los estudiantes evaluados en mayor medida logran llegar a capacidades que involucran resolver problemas de sumas o restas, o resolver ejercicios de fracciones. Sin embargo, un aspecto interesante se aprecia en la parte inferior donde en ambas muestras el porcentaje de evaluados que no cuentan con habilidades numéricas es similar, siendo las mayores diferencias al interior del continuo de capacidades evaluadas.

Cuadro 12. Distribución de estudiantes evaluados por puntaje en la prueba MIA+

	Jurisdicción A			Jurisdicción C		
	N	% relativo	% acumulado	N	% relativo	% acumulado
Resuelve fracciones	84	2.4	2.4	455	14.1	14.1
Resuelve problemas	77	2.2	4.7	246	7.6	21.6
Resuelve divisiones	141	4.1	8.8	229	7.1	28.7
Resta llevando	666	19.4	28.1	823	25.4	54.1
Resta sin llevar	403	11.7	39.9	316	9.8	63.9
Suma 2 dígitos	543	15.8	55.6	303	9.4	73.3
Suma 1 dígito	761	22.1	77.8	409	12.6	85.9
Decodifica números	475	13.8	91.6	247	7.6	93.5
No sabe leer números	290	8.4		210	6.5	

Fuente: Bases de datos de línea de base 2021

Ilustración 1. Distribución de densidad de los puntajes del MIA+



Fuente: Bases de datos de línea de base 2021

3.2.3. Validez predictiva

Junto con la validez de constructo, de manera complementaria se realizaron algunos ejercicios para identificar la validez predictiva. En este sentido, en primer lugar se procedió a estimar la correlación del puntaje estimado para la escala SMS y variables demográficas como el sexo,

edad y nivel socioeconómico. No se encontraron diferencias por sexo, mientras que en el caso de la edad y nivel socioeconómico se encontraron asociaciones positivas y significativas.

Cuadro 13. Correlaciones entre el puntaje de la escala SMS y variables demográficas

	Jurisdicción A	Jurisdicción B	Jurisdicción C
Mujer	-0.022 (0.196)	0.015 (0.399)	0.024 (0.174)
Edad	0.161 (0.000)	0.195 (0.000)	0.195 (0.000)
Nivel Socioeconómico	0.107 (0.000)	0.042 (0.017)	0.138 (0.000)

Nota: probabilidad de que la correlación tome el valor de 0 entre paréntesis.

Fuente: Elaboración propia a partir de las bases de datos de línea de base 2021.

De igual forma, se estimó la correlación entre el puntaje de la escala MIA+ y las variables demográficas. Se puede apreciar en el siguiente cuadro hay una correlación positiva y significativa con la edad y el nivel socioeconómico de las familias, y no existiría diferencias por sexo entre los estudiantes evaluados.

Cuadro 14. Correlaciones entre el puntaje de la escala MIA+ y variables demográficas

	Jurisdicción A	Jurisdicción C
Mujer	0.016 (0.356)	0.017 (0.326)
Edad	0.158 (0.000)	0.173 (0.000)
Nivel Socioeconómico	0.085 (0.000)	0.091 (0.000)

Nota: probabilidad de que la correlación tome el valor de 0 entre paréntesis.

Fuente: Elaboración propia a partir de las bases de datos de línea de base 2021.

En cuanto a la correlación entre el puntaje de la escala SMS y la escala MIA+, en el siguiente cuadro se aprecia que esta es positiva y significativa, estando los índices de correlación por encima de 0.50. De esta manera, se puede apreciar que tanto el puntaje del SMS como del

MIA+ están apuntando en la misma dirección y estarían midiendo las habilidades matemáticas de los y las estudiantes.

Cuadro 14. Coeficientes de correlación entre el puntaje de la escala MIA+ y la escala SMS

	Índice
Jurisdicción A	0.60 (0.000)
Jurisdicción B[1/]	0.71 (0.000)
Jurisdicción C	0.52 (0.000)

1/ En el caso de esta jurisdicción, el puntaje del MIA+ fue generado mediante la suma simple de todas las respuestas correctas que tuvo cada estudiante a los ítems de la escala.

Nota: probabilidad de que la correlación tome el valor de 0 entre paréntesis.

Fuente: Elaboración propia a partir de las bases de datos de línea de base 2021

Esto se confirma por medio de una regresión lineal simple para comprobar si el SMS predecía significativamente MIA+ en el caso de la jurisdicción A. El modelo de regresión ajustado fue: .0554 + -0.007xMIA+

La regresión global fue estadísticamente significativa (R2 = 0.30, F(1, 3438)= 1500.16, p = 0.0000). De esta manera, se aprecia una relación entre SMS y MIA+ (0.554, sig al 1%).

3.3 Análisis TRI del instrumento SMS

Dados los resultados obtenidos en la sección previa, el módulo SMS confirmó su unidimensionalidad y, por ende, en él se puede usar un modelo TRI para poder estimar las habilidades matemáticas de los y las estudiantes evaluados.

Cabe señalar que los ítems administrados en el módulo SMS son de respuesta abierta, motivo por el cual no se considera como opción el modelo de tres parámetros de la Teoría de Respuesta al Ítem. Así, se realiza la prueba de ratio de verosimilitud *(Likelihood Ratio test*) para determinar si es mejor modelar usando un modelo 1PL o 2PL. Los resultados se aprecian en el siguiente cuadro donde se indica que el mejor modelo es el de dos parámetros en todas las muestras, es decir, que no solo la dificultad del ítem determina la habilidad de los y las estudiantes, sino también la discriminación de cada ítem.

Cuadro 15. Prueba de ratio de verosimilitud entre el modelo de 1 y 2 parámetros

	Loglikelihood	Grados de libertad	AIC	BIC
Jurisdicción A				
1pl	14059.61	9	28137.22	28192.51
2pl	-13691.95	16	27415.90	27514.19
Jurisdicción B				
1pl	-14344.47	9	28706.94	28761.60
2pl	-14151.29	16	28334.58	28431.77
Jurisdicción C				
1pl	-14387.80	9	28793.59	28848.34
2pl	-14268.74	16	28569.48	28666.80

Fuente: Elaboración propia a partir de las bases de datos de línea de base 2021.

Una vez definido el modelo, se procedió a estimar el modelo TRI de 2 parámetros. En el siguiente cuadro se aprecian los indicadores de dificultad y discriminación de cada uno de los ítems. Se aprecia altos niveles de discriminación de los ítems, es decir, permite discriminar bien entre aquellos y aquellas estudiantes de menor y mayor niveles de habilidad. Mientras, en el caso de la dificultad de los ítems, se puede apreciar que los niveles de dificultad están por encima del valor promedio de la escala (media: 0), siendo el ítem con mayor nivel de dificultad el de razonamiento lógico (math_sms7), que es un aspecto que podría ser abordado en futuras investigaciones. Por otro lado, se puede apreciar que no hay mucha variación en el ordenamiento de los ítems para las diferentes muestras de estudio, por lo que mostraría la consistencia de la prueba en diferentes contextos.

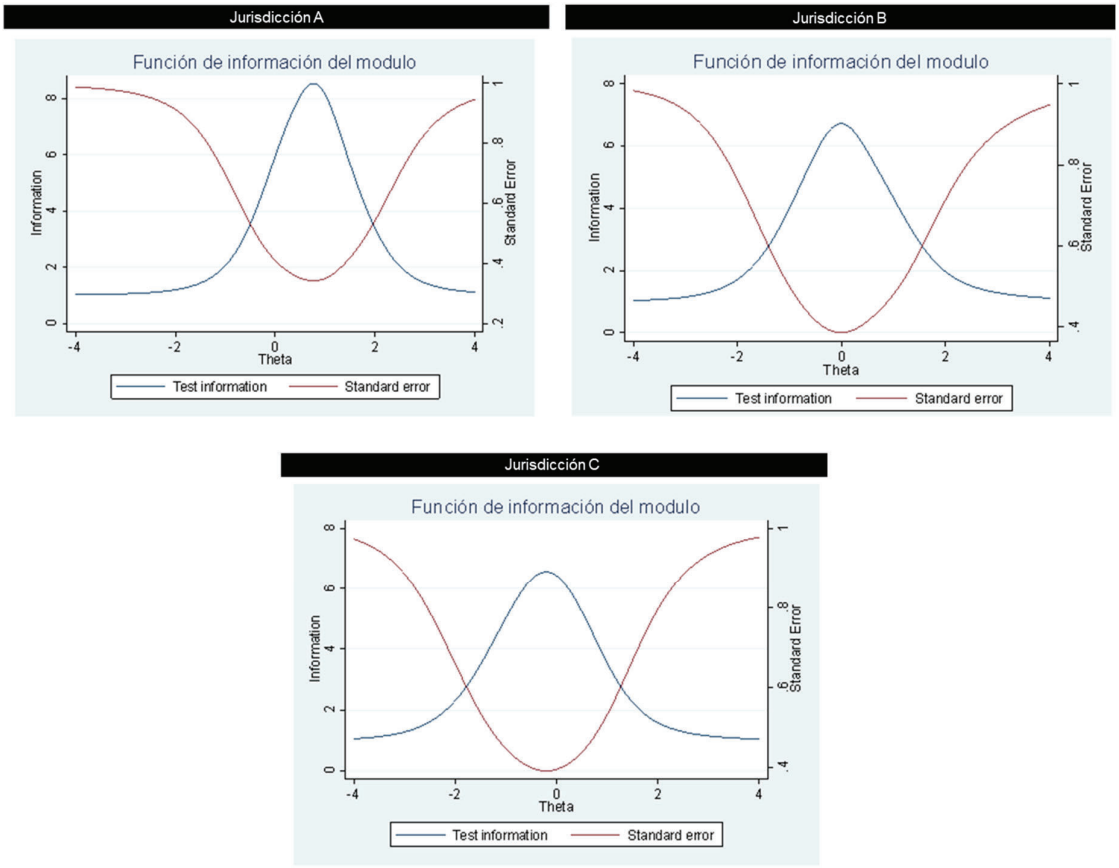
Cuadro 16. Índices de discriminación y dificultad de cada ítem

Jurisdicción A		Jurisdicción B		Jurisdicción C	
Ítems	Valor	Ítems	Valor	Ítems	Valor
Discriminación					
math_sms2	2.626	math_sms2	2.038	math_sms2	1.785
math_sms6	2.155	math_sms6	1.981	math_sms6	1.864
math_sms3	2.846	math_sms4	2.728	math_sms1	1.718
math_sms1	1.368	math_sms10	1.126	math_sms3	2.226
math_sms10	1.197	math_sms3	2.115	math_sms10	1.229
math_sms4	3.071	math_sms1	1.188	math_sms4	2.299
math_sms7	0.610	math_sms5	2.334	math_sms5	2.317
math_sms5	2.503	math_sms7	0.842	math_sms7	1.014
Dificultad					
math_sms2	0.128	math_sms2	-0.702	math_sms2	-0.977
math_sms6	0.169	math_sms6	-0.450	math_sms6	-0.938
math_sms3	0.713	math_sms4	-0.004	math_sms1	-0.491
math_sms1	0.786	math_sms10	0.011	math_sms3	-0.274
math_sms10	0.908	math_sms3	0.114	math_sms10	-0.096
math_sms4	0.935	math_sms1	0.424	math_sms4	-0.072
math_sms7	0.998	math_sms5	0.893	math_sms5	0.552
math_sms5	1.536	math_sms7	1.619	math_sms7	0.597

Fuente: Elaboración propia a partir de las bases de datos de línea de base 2021.

Otro aspecto que es clave revisar es la relación entre la información que proporcionan en conjunto los ítems usados y el margen de error que se tienen en la medición de las habilidades. Es así que el siguiente gráfico muestra la relación entre la función de información de la escala o el módulo SMS y el error en la estimación de la habilidad. Se considera que se está midiendo bien la habilidad de los y las estudiantes si la curva de información (línea azul) está por encima de la curva del error de medición de las habilidades (línea roja). En el caso de la jurisdicción A, se aprecia que la prueba es buena o adecuada para medir las habilidades de los y las estudiantes que se ubican en la habilidad promedio o por encima de esta, mientras a niveles de habilidad por debajo del promedio, el margen de error es mucho mayor. Este aspecto guarda relación con el hecho de que no se cuenta con ítems medianamente fáciles para evaluar la capacidad matemática de los y las estudiantes. Una imagen distinta se tiene para la muestra de las jurisdicciones B y C donde la prueba resulta adecuada para medir la habilidad de los estudiantes tanto por encima como por debajo del promedio de habilidad, mostrando de esta manera que se cuenta con ítems con un buen nivel de información.

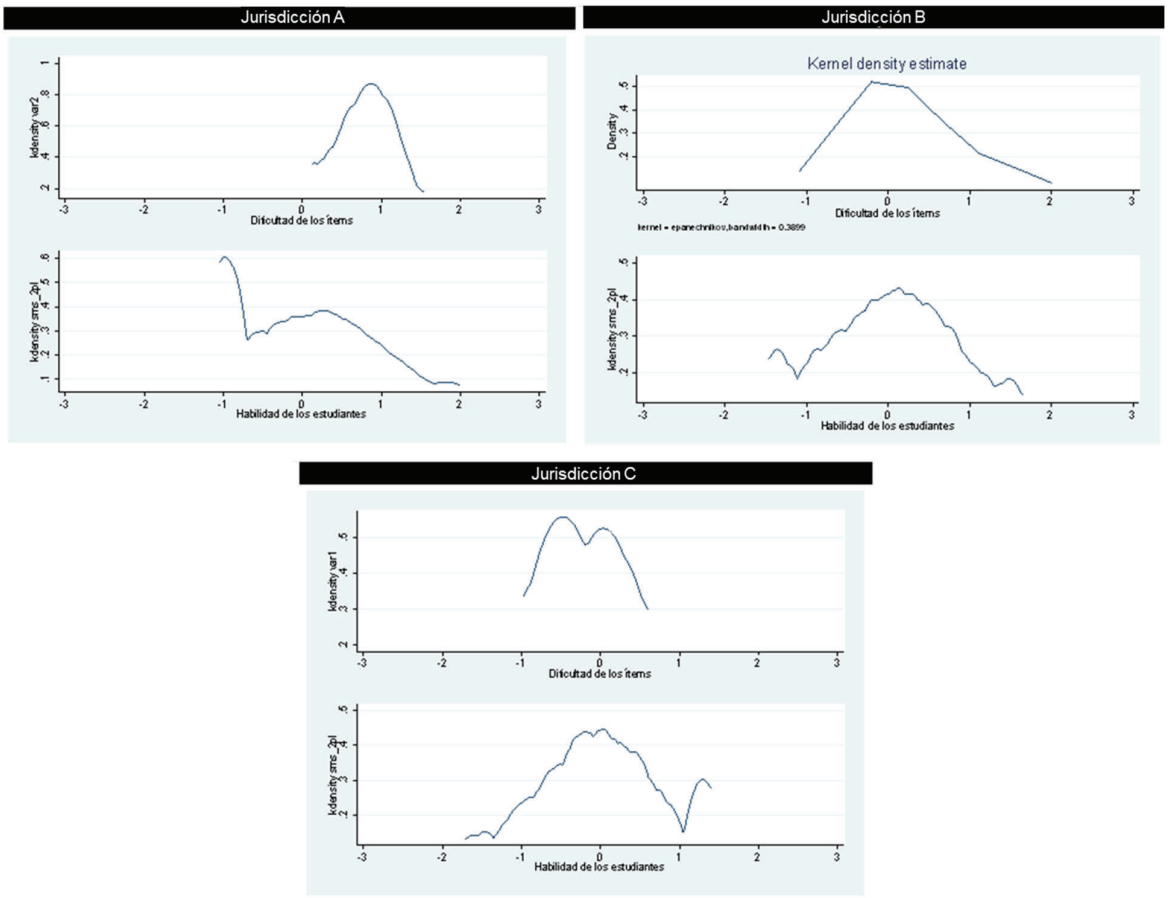
Ilustración 2 Curva de información de la escala y curva de error de medición para los diferentes niveles de habilidad



Fuente: Elaboración propia a partir de las bases de datos de línea de base 2021.

Una de las ventajas del modelo TRI es que permite poner en una misma escala la habilidad de los y las estudiantes y la dificultad de los ítems, motivo por el cual, se procedió a graficar la distribución de los puntajes de los estudiantes y la distribución de la dificultad de los ítems (ver gráfico 4). En el caso de la jurisdicción A, se aprecia que en promedio los ítems han sido difíciles para los y las estudiantes evaluados, siendo baja la proporción de estudiantes que pueden responder todos los ítems del módulo SMS. Mientras que para los y las estudiantes de la jurisdicción B y la jurisdicción C, se aprecia una superposición de ambas curvas con lo que se apreciaría que los diferentes ítems de la prueba pueden ser respondidos por los y las estudiantes.

Ilustración 3. Distribución de la habilidad de los y las estudiantes y la dificultad de los ítems



Fuente: Elaboración propia a partir de las bases de datos de línea de base 2021.

3.4. Nivel adecuado para cada estudiante

Uno de los principales aportes del Proyecto es la posibilidad de que las tutorías respondan a las necesidades singulares de cada estudiante, interviniendo oportunamente en función de los resultados de la evaluación diagnóstica inicial y del monitoreo que se realiza a lo largo del periodo de trabajo. Es por ello que se consideró necesario analizar en qué medida cada uno de los instrumentos podría identificar con precisión el nivel adecuado los estudiantes.

SMS

En el caso del instrumento SMS, los reactivos no se encuentran ordenados por nivel de dificultad. Por lo que, para esta investigación, nos dimos a la tarea de ordenarlos por nivel de dificultad. En este sentido, se identificaron los niveles de suma, resta, multiplicación, división y fracciones.

- El nivel suma se definió como aquellos niños que no pueden realizar una suma, independientemente de que puedan realizar otras operaciones matemáticas.

- El nivel resta se definió como aquellos niños que pueden sumar pero que no pueden resolver una resta, independientemente de que puedan realizar otras operaciones.
- El nivel multiplicación se definió como aquellos niños que pueden restar pero que no pueden resolver una multiplicación, independientemente de que puedan realizar otras operaciones.
- El nivel división se definió como aquel nivel donde los niños pueden realizar multiplicaciones, pero no pueden dividir, independientemente de que puedan realizar otras operaciones.
- Y el nivel fracciones se identificó como aquellos niños que pudieron dividir y multiplicar.

En esta investigación se consideró identificar en el nivel adecuado asumiendo que tuvieran un cero, es decir que no hubieran podido responder el reactivo, se consideraría el nivel en que tendrían que cursar la intervención. Así, por ejemplo, un niño que respondió inadecuadamente a la suma pero adecuadamente a la resta, se quedaría en el nivel de suma. De esta forma, utilizando solo el instrumento SMS el nivel adecuado de los sujetos sería el siguiente:

Cuadro 17. Ubicación nivel adecuado usando escala SMS

	Jurisdicción A		Jurisdicción B		Jurisdicción C	
	N	%	N	%	N	%
Suma	1857	54%	945	29%	777	24%
Resta	771	22%	914	28%	752	23%
Multiplicación	357	10%	315	10%	480	15%
División	248	7%	517	16%	502	16%
Fracciones	207	6%	519	16%	727	22%
Total	3440	100%	3210	100%	3238	100%

MIA+

- En el caso del instrumento MIA+, si identificó el nivel suma, resta, multiplicación y fracciones.
- El nivel suma agrega aquellos niños que no identifican números y que no son capaces de realizar una suma,
- El nivel resta identifica niños que pueden sumar pero que no pueden restar,
- El nivel multiplicación identifica niños que pueden restar pero que no pueden dividir puntos,
- El nivel fracciones identifica niños que si pueden dividir.

En el caso del instrumento MIA+, al no tener un reactivo específico sobre multiplicación, se asume que los niños que pueden restar, pero no pueden dividir, necesitan comenzar por el nivel multiplicación.

Cuadro 18. Ubicación nivel adecuado utilizando MIA+

MIA+	Jurisdicción A		Jurisdicción B		Jurisdicción C	
	N	%	N	%	N	%
Suma	1526	44%	439	14%	866	27%
Resta	946	28%	651	20%	619	19%
Multiplicación	667	19%	557	17%	823	25%
Fracciones	301	9%	1563	49%	930	29%
Total	3440	100%	3210	100%	3238	100%

Al hacer una comparación entre los niveles identificados por SMS y por MIA+ encontramos algunas diferencias en los resultados que podrían inducir a sesgos en la identificación adecuada de cada nivel. Por ejemplo, el siguiente cuadro muestra la distribución entre los niveles según SMS y MIA+ para la jurisdicción A. Aquí es posible ver que, si bien 1180 sujetos coinciden en el nivel "suma" según ambas escalas, hay 58 sujetos que estarían en nivel "fracciones" en MIA+ pero "suma" en SMS.

Cuadro 20. Comparación ubicación según escala utilizada, caso Jurisdicción A.

Ubicación MIA+	Ubicación SMS					Total
	Suma	Resta	Multiplicación	División	Fracciones	
Suma	1180	243	63	28	12	1526
Resta	431	315	115	57	28	946
Multiplicación	196	155	143	123	50	667
Fracciones	58	58	36	40	117	301
Total	1857	771	357	248	207	3440

Nota: Prueba Pearson χ^2 = 1.4e+03 (p=0.000)

Así, si bien ambos instrumentos tienen una correlación alta, este ejercicio permitió ver diversas observaciones con inconsistencias importantes que habría que eliminar para disminuir los sesgos. En algunos casos tienen calificaciones muy altas en el instrumento SMS y muy bajas y MIA+ y viceversa. Por ello, se fijó un criterio ex post que permitiera eliminar las inconsistencias del grupo de tratamiento para no incluir sesgos, a saber: se optó por identificar al sujeto en el nivel más bajo que hubiera obtenido en las dos opciones, asumiendo uno de los principios del TaRL, que tiene que ver con enseñar en el nivel adecuado, donde el sujeto realmente se siente capaz de resolver los desafíos que se le presentan. Para ello, se creó un nivel que identificó el nivel más bajo entre ambos instrumentos.

Cuadro 19. Ubicación nivel adecuado utilizando criterio de nivel más bajo.

	Jurisdicción A		Jurisdicción B		Jurisdicción C	
Suma	2203	64%	1047	33%	1230	38%
Resta	728	21%	962	30%	779	24%
Multiplicación	352	10%	365	11%	622	19%
División	40	1%	344	11%	156	5%
Fracciones	117	3%	492	15%	451	14%
Total	3440	100%	3210	100%	3238	100%

Al analizar la correlación entre los niveles más bajos identificados y los instrumentos, hubo como se esperaba correlaciones altas, aunque con diferencias importantes entre las jurisdicciones.

Cuadro 20.Corelaciones entre ubicación del nivel más bajo y resultados de instrumentos

Escala Jurisdicción	MIA	SMS
Jurisdicción A	0.73***	0.83***
Jurisdicción B	0.64***	0.94***
Jurisdicción C	0.78***	0.76***

*** p<0.001

Así, para identificar el nivel "más bajo" ambos instrumentos parecen indicados. En el caso del instrumento SMS, si bien parece identificar mejor los niveles más bajos, hay que observar, por un lado, que solo utiliza una operación para identificar el acierto/error, por lo que puede sesgar los resultados y, por otro, que requiere ordenar los ítems según nivel de dificultad. En el caso de MIA+, por otro lado, permite identificar de manera más intuitiva los niveles reales de aprendizajes (por la forma de ordenamiento de más fácil a más difícil que tiene el instrumento), y puede identificar un rango más amplio de dificultad, lo que resulta útil para no subestimar o sobreestimar los niveles de los aprendizajes.

CONCLUSIONES Y RECOMENDACIONES

El presente documento presenta una revisión a la base de datos de las jurisdicciones A, B y C que cuentan con información de las escalas MIA+ y SMS. Ambas escalas miden las habilidades matemáticas fundamentales de los y las estudiantes entre 9 y 14 años. Los análisis presentados hasta acá han tenido como finalidad poder evaluar la confiabilidad y la evidencia de validez de constructo para cada una de las escalas de forma separada.

Para cada una de las escalas se evaluó la confiabilidad mediante los índices de *Cronbach* y *Omega*, mientras que en el caso de la validez, se procedió a usar el análisis factorial confirmatorio para validar la existencia de un solo factor latente en la escala SMS, y que esta refleje las habilidades numéricas de los y las estudiantes evaluados.

A continuación, se sistematizan los distintos resultados para cada uno de los instrumentos analizados, identificando aquellos aspectos que resultan favorables en cada caso:

Cuadro 21. Características de confiabilidad y validez de los instrumentos para medir aprendizajes

Atributos	SMS	MIA+
Utilidad (principal propósito del instrumento)	Estimar el efecto de la tutoría remota y comparar los resultados obtenidos en América Latina con otros países que han utilizado el modelo de tutorías telefónicas (como Botsuana, India o Nepal).	Indicar el nivel de avance de los estudiantes en distintos ejes de contenido
Discriminación de los ítems	Adecuados, con correlaciones ítem-resto de la prueba por encima de 0.20.	Adecuados, con correlaciones ítem-resto de la prueba mayores al 0.20.
Confiabilidad	Adecuados niveles de confiabilidad (≥0.70)	Adecuados niveles de confiabilidad (≥0.70)
Confiabilidad – sub-análisis	Tomando los 6 primeros ítems, adecuado (≥0.70)	Tomando los 5 primeros ítems, adecuado (≥0.70)
Unidimensionalidad	Con solo una dimensión latente ya sea para la escala completa como para la versión reducida de la misma.	No aplica
Independencia local	Sí	No
Regla de discontinuación	No	Sí

Dificultad de los ítems	Los ítems no han sido medianamente difíciles (en promedio). Los ítems no se encuentran ordenados por dificultad.	Los ítems muestran la progresión de dificultad esperada de acuerdo al tipo de habilidad que se mide.
Escala de los puntajes	Puntajes en escala continua.	Puntajes en escala continua.

Con relación a la confiabilidad de los instrumentos utilizados, los resultados psicométricos nos muestran que **ambas escalas (MIA+ y SMS) cuentan con adecuados niveles de confiabilidad (≥0.70) para ambas muestras de estudio**. Luego, al realizar el análisis de validez de la escala SMS, tanto para las jurisdicciones A y C, se pudo apreciar que estas cuentan con solo una dimensión latente ya sea para la escala completa como para la versión reducida de la misma.

Respecto de las evidencias que arroja cada uno de los instrumentos, se observa que **ambos instrumentos permiten contar con evidencias sobre el efecto de las tutorías remotas**. En el caso del SMS, al tratarse del mismo instrumento utilizado en otros casos, posibilita a su vez la comparación de los resultados obtenidos en América Latina con otros países que han utilizado el modelo de tutorías telefónicas (como Botsuana, India o Nepal). En el caso de **MIA+, también resulta posible indicar el nivel de avance de los estudiantes en distintos ejes de contenido**, de modo de propiciar intervenciones oportunas.

En relación a la validez predictiva de los instrumentos, la correlación del puntaje estimado para la escala SMS y variables demográficas como el sexo, edad y nivel socioeconómico **no arroja diferencias por sexo, mientras en el caso de la edad y nivel socioeconómico se encontraron asociaciones positivas y significativas.**

En cuanto a la correlación entre el puntaje de la escala SMS y la escala MIA+, se aprecia que esta es positiva y significativa, estando los índices de correlación por encima de 0.50. De esta manera, se puede apreciar que tanto **el puntaje del SMS como del MIA+ están apuntando en la misma dirección y estarían midiendo las habilidades matemáticas de los y las estudiantes.**

Para poder medir el efecto de las tutorías, el instrumento SMS tiene como sus principales ventajas sus propiedades psicométricas y la facilidad de su aplicación. Sus desventajas tienen que ver con cierto sesgo implícito al tener una sola opción de respuesta por cada reactivo, y con un rango relativamente corto en el rango de dificultad que mide.

De igual forma, el instrumento MIA+ tiene como principales ventajas la progresión en la complejidad de los factores que mide, el rango más amplio de dificultad que mide y la simpleza de su interpretación y aplicación. Sus desventajas se encuentran en el salto de complejidad existente entre los campos aditivos y multiplicativos, donde requiere más ítems para identificar con mayor exactitud las progresiones.

Para futuras mediciones existen tres opciones: decidir la utilización de una de estas escalas, de ambas, o bien construir un tercer instrumento que permita integrar los puntos fuertes que posee cada instrumento.

La primera opción, la selección de uno de los instrumentos puede estar en sus características de unifactorialidad (SMS) o multi-factorialidad (MIA+). El instrumento SMS es óptimo para dar resultados generales sobre las habilidades matemáticas en general: su aplicación es muy rápida y sus resultados permiten tener una medida para poder identificar factores asociados a los aprendizajes de matemáticas (incluyendo el efecto de intervenciones educativas). El instrumento MIA+ puede hacer esto, además de dar cuenta de los niveles de progresión de los estudiantes

La segunda opción, aplicar ambos instrumentos, resulta posible aunque un tanto redundante: tanto la medición más general sobre habilidades matemáticas como la identificación de los niveles de aprendizajes se pueden hacer con uno u otro instrumento.

Por ello, la tercera opción resulta relevante: poder crear un solo instrumento, que mantenga su simpleza, validez y confiabilidad, que sea fácil de aplicar e interpretar y que pueda cumplir con los objetivos que se persiguen: medir los aprendizajes fundamentales a la vez de identificar el nivel adecuado en estudiantes.

Bajo esta lógica, creemos que la tercera opción podría basarse en el instrumento MIA+, considerando la utilidad de este instrumento para proporcionar información desagregada por eje de contenido. Este podría pensarse como un instrumento diferente que contenga las siguientes características:

4.1.1. Definir con mayor claridad el objetivo de la evaluación de aprendizajes.

La evaluación realizada debe orientar la intervención didáctica de las tutorías remotas, posibilitando conocer el nivel adecuado de cada estudiante a lo largo del proceso. A su vez, debe ser capaz de proporcionar información nominal de cada estudiante, de modo de acompañar la mejora en los aprendizajes, incluso una vez finalizado el proyecto.

En este sentido, el nuevo instrumento tiene que definir de manera clara los dos objetivos que el instrumento tiene: medir hasta qué capacidad numérica son capaces de llegar los y las estudiantes evaluados, así como identificar en una escala válida y confiable aprendizajes fundamentales de matemáticas.

La discontinuidad de la prueba MIA+, diseñada para que un o una estudiante no pase a una pregunta adicional si no responde una pregunta previa, hace posible que los ítems de la escala sean dependientes y por ende no sea necesario ver la validez de constructo dado que el puntaje que se obtiene no refleja un rasgo latente sino es un puntaje ordinal donde cada valor indica tiene un carácter acumulativo. El análisis de constructo de MIA+ (ver 3.2.) muestra la existencia de tres factores, algunos con muy pocos reactivos. De ahí que el siguiente instrumento pudiera considerar la creación de más reactivos por nivel de

tal forma que pueda medirse la unifactorialidad de cada uno de los niveles con procedimientos como TRI, y con eso solventar tanto las limitaciones del instrumento SMS (el sesgo de aplicar un solo reactivo) como del instrumento MIA+ (no tener una validez de constructo por factor).

4.1.2. Incorporar mayor cantidad de ítems específicos.

A partir del análisis realizado, resulta necesario revisar la escala MIA+ e incluir ítems que permitan medir mejor el continuo de habilidades matemáticas de los y las estudiantes, y de esta forma evitar el salto en la dificultad de los ítems (tasa de acierto) como la observada tanto en la jurisdicción A como en la C entre los ejercicios que miden las habilidades de restas y la división (ver cuadro 23).

Cuadro 22. Índices de dificultad de los ítems del módulo MIA+

Aspecto		Jurisdicción A	Jurisdicción C
Lectura de números	math_plus1	0.916	0.935
Suma sin llevar	math_plus2	0.777	0.859
Suma llevando	math_plus3	0.556	0.733
Resta sin prestar	math_plus4	0.399	0.639
Resta prestando	math_plus5	0.281	0.541
Multiplicaciones	SMS u otra fuente	Salto de más de 15 pp	
Divisiones	math_plus6	0.088	0.287
Resolución de problemas	math_plus7	0.047	0.216
Resolución de problemas	math_plus8	0.033	0.170
Fracciones	math_plus9	0.025	0.141

Fuente: Bases de datos de línea de base 2021

Elaboración propia

Dado lo anterior, se recomienda incluir ítems similares al administrado en la escala SMS dado que cumple con la progresión esperada de dificultad que plantea el MIA+ para las jurisdicciones A y C (ver cuadro 21).

Cuadro 23. Ítems que se pueden incorporar en la escala MIA+

Ítem	Tasa de acierto	Fuente
¿Cuánto es 28 multiplicado por 3?	0.52 (Jurisdicción C) 0.21 (Jurisdicción A)	Escala SMS
23 x 19 =	0.41 (Promedio Internacional) 0.23 (Chile)	TIMSS[1/] 2011 4to grado (ID: M051203)
27 x 43 =	0.51 (Promedio Internacional) 0.25 (Chile) 0.44 (Buenos Aires, Argentina)	TIMSS 2015 / 2019 4to grado (ID: M061273)
6 x 312 =	0.64 (Promedio Internacional) 0.44 (Chile) 0.65 (Buenos Aires, Argentina)	TIMSS 2015 /2019 4to grado (ID: M061271)

1/ TIMSS es desarrollado por el *International Association for the Evaluation of Educational Achievement* (IEA) desde 1995 a estudiantes de 4to y 8vo grado.

4.1.3. Incluir ítems adicionales de menor nivel de dificultad

Finalmente, en relación a la escala del SMS, se ha encontrado que cuenta con solo una dimensión, aspecto que permitió estimar sus puntajes usando el modelo de TRI (dos parámetros) encontrándose que los ítems permiten medir bien las habilidades pero de aquellos y aquellas estudiantes que tienen niveles de habilidad promedio o por encima de este en la jurisdicción A; mientras que en la jurisdicción C mostró tener un buen comportamiento la escala y medir bien la habilidad de los y las estudiantes a los largo de los diferentes niveles de habilidad. Motivo por el cual, se recomienda incluir ítems adicionales de menor nivel de dificultad (o más sencillos) para poder capturar mejor la habilidad de los y las estudiantes evaluados en todo el continuo.

Para concluir, se observa que los instrumentos utilizados cuentan con evidencias de confiabilidad y validez necesarias para poder proporcionar las evidencias de aprendizajes requeridas en el marco del proyecto de tutorías remotas. Se considera que el instrumento MIA+ resulta más adecuado para tomar de base para enriquecer la propuesta de evaluación, priorizando el objetivo de la intervención de identificar los niveles adecuados de cada estudiante.

Se espera que la versión mejorada del instrumento sea piloteada a la brevedad, posibilitando medir de manera simple pero aún más robusta los efectos de estas tutorías en los aprendizajes fundamentales de matemática.

En conclusión, estos instrumentos permiten la evaluación rigurosa de una intervención como la de tutorías remotas. Con algunas mejoras podrían constituir una base para evaluar otros programas e intervenciones de forma confiable y costo-efectiva. En el futuro esta nueva herramienta podría estar disponible para los equipos y los gobiernos que deseen evaluar de forma rápida y costo efectiva si sus programas o políticas van encaminados a lograr los resultados esperados. Al contar con un número limitado de ítems y poder aplicarse de forma relativamente rápida, estos instrumentos podrían ser combinados con otros o incluso aplicados más de forma recurrente. Además, al tratarse de un instrumento de fácil aplicación, ya que no depende de dispositivos y/o conexión a internet de banda ancha, resulta posible su utilización en la población más vulnerable. Por último, al tratarse de ítems relativamente sencillos no sería costoso modificarlos para que continúen midiendo las habilidades matemáticas mitigando el riesgo de que los estudiantes sean preparados para responder estos instrumentos tal y como se presentan aquí de forma idéntica, por ejemplo aprendiendo de memoria las repuestas. Al ser aplicados de forma individual también se mitiga el riesgo de que las repuestas de los estudiantes no sean auténticas ya sea por que reciban ayuda de otros o por cualquier otro motivo.

REFERENCIAS

Angrist, N., Bergman, P., Evans, D. K., Hares, S., Jukes, M. C. H. y Letsomo, T. (2020). Practical Lessons for Phone-Based Assessments of Learning. *BMJ Global Health, 5*(7): e003030. https://doi.org/10.1136/bmjgh-2020-003030.

Angrist, N., Bergman, P., Brewster, C., y Matsheng, M. (2020). Stemming learning loss during the pandemic: A rapid randomized trial of a low-tech intervention in Botswana. Available at SSRN 3663098. https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=3663098

Banerjee, A., Banerji, R., Berry, J., Duflo, E. , Kannan, H., Mukerji, S., Shotland, M. y Walton, M. (2016). Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of "Teaching at the Right Level" in India. *National Bureau of Economic Research (NBER) Working Paper No. 22746*.

Cho, E., y Kim, S. (2015). Cronbach's coefficient alpha: Well known but poorly understood. *Organizational research methods, 18*(2), 207-230.

Crocker, L., y Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt.

Egon, S. (1931). The Test of Significance for the Correlation Coefficient. *Journal of the American Statistical Association, 26* (174): 128–34. https: //doi.org/10.1080/01621459.1931.10503208.

Fox, J. (2016). *Applied Regression Analysis and Generalized Linear Models*. SAGE.

Hair, J.E., Anderson, R.E. Tatham, R.L. y Black, W.C. (2006). *Multivariate Data Analysis*. Prentice Hall College.

Hevia, F. J. y Vergara-Lope, S. (2016). Evaluaciones Educativas Realizadas Por Ciudadanos En México: Validación de La Medición Independiente de Aprendizajes. *Innovación Educativa 16* (70): 85–110. http://www.openaccess.innovacion.ipn.mx/index.php/inovacion/article/ viewFile/53/38.

Howell, D. C. (2006). *Statistical Methods for Psychology.* Wadsworth Publishing.

Kaiser, H. (1970). A second generation Little Jiffy. *Psychometrika, 35,* 401-15.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum Associates, Inc.

McDonald, R. P. (2013). *Test Theory: A Unified Treatment*. Psychology Press. Pearson.

Nunnally, J. C. y Bernstein, I. H., (1994). *Psychometric theory.* McGraw-Hill.

Oliva, T. A., Oliver, R. L. y MacMillan, I. C. (1992). A catastrophe model for developing service satisfaction strategies. *Journal of Marketing, 56,* 83-95.

Reckase, M. D. (1979) Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*, 207- 230.

Tristán, A. (2013). *Análisis de Rasch Para Todos. Una Guía Simplificada Para Evaluadores Educativos.* IEIA.

Zwick, R. (1987). Assessing the dimensionality of NAEP reading data. *Journal of Educational Measurement, 24*(4), 293-308.

ANEXOS

Anexo 1

#	Preguntas	Respuesta correcta		Respuesta incorrecta	
1	**Juan tiene 47 manzanas y las organiza por La Tabla de Valores. ¿Cuántas DECENAS tiene?**	Responde 4 decenas en menos de 30 segundos	Sí (1)	No responde, responde otro número o demora más de 30 segundos	No (0)
2	**¿Cuánto es 52 más 39?**	Responde 91 en menos de 2 minutos	Sí (1)	No responde, responde otro número o demora más de 2 minutos	No (0)
3	**¿Cuánto es 42 menos 29?**	Responde 13 en menos de 2 minutos	Sí (1)	No responde, responde otro número o demora más de 2 minutos	No (0)
4	**¿Cuánto es 28 multiplicado por 3?**	Responde 84 en menos de 2 minutos	Sí (1)	No responde, responde otro número o demora más de 2 minutos	No (0)
5	**¿Cuánto es 65 dividido por 8?**	Responde 8.1 ó 8.12 en menos de 2 minutos	Sí (1)	No responde, responde otro número o demora más de 2 minutos	No (0)
6	**Un hombre conduce 24 km. Luego conduce 17 km. ¿Cuántos kilómetros ha recorrido en total?**	Responde 41 en menos de 2 minutos	Sí (1)	No responde, responde otro número o demora más de 2 minutos	No (0)
7	**El día anterior de pasado mañana es sábado. ¿Qué día es hoy?**	Responde "viernes" en menos de 2 minutos	Sí (1)	No responde, responde otro día de la semana o demora más de 2 minutos	No (0)
8	**El encuestador dice "Quieres intentar responder a otra pregunta de lógica"**	Pasa pregunta 8	Sí (1)	Pasa pregunta 9	No (0)
9	Sí responde si, el encuestador lee: **El día después de ayer es martes. ¿Qué día es hoy?**	Responde "martes" en menos de 2 minutos	Sí (1)	No responde, responde otro día de la semana o demora más de 2 minutos	No (0)
10	**¿Cuánto es 3/8 + 4/8?**	Responde 7/8 en menos de 2 minutos	Sí (1)	No responde, responde otro número o demora más de 2 minutos	No (0)

Anexo 2

	Instrucción	Respuesta correcta		Respuesta incorrecta	
1	***Elige dos números y léelos en voz alta:*** *Si se equivoca en uno, dar la opción de que elija un tercer número*	Identifica correctamente al menos dos de tres números elegidos hasta en dos intentos	Sí (1)	No identifica correctamente al menos dos de tres números elegidos hasta en dos intentos	No (0)
2	***Elige dos sumas y resuélvelas*** *Si se equivoca en una, dar la opción de que elija una tercera suma*	Resuelve correctamente al menos dos de tres sumas elegidas hasta en dos intentos 3+8=11 9+6=15 8+4=12 9+2=11 5+9=14 5+8=13	Sí (1)	No resuelve correctamente al menos dos de tres sumas elegidas en dos intentos.	No (0)
3	***Elige dos sumas y resuélvelas*** *Si se equivoca en una, dar una la opción de que elija una tercera suma*	Resuelve correctamente al menos dos de tres sumas elegidas hasta en dos intentos 46+28=74 27+57=84 34+18=52 36+48=84 61+29=90 19+72=91	Sí (1)	No resuelve correctamente al menos dos de tres sumas elegidas en dos intentos.	No (0)
4	***Elige dos restas y resuélvelas***	Resuelve correctamente al menos dos de tres restas elegidas hasta en dos intentos	Sí (1)	No resuelve correctamente al menos dos de tres restas elegidas en dos intentos.	No (0)

	Si se equivoca en una, dar una la opción de que elija una tercera resta	99-76=23 67-43=24 53-11=42 44-13=31 54-23=31 97-65=32			
5	***Elige dos restas y resuélvelas*** *Si se equivoca en una, dar una la opción de que elija una tercera resta*	Resuelve correctamente al menos dos de tres restas elegidas hasta en dos intentos 74-35=39 34-17=17 31-14=17 48-29=19 93-44=49 77-18=59	Sí (1)	No resuelve correctamente al menos dos de tres restas elegidas en dos intentos.	No (0)
6	***Elige dos divisiones y resuélvelas*** *Si se equivoca en una, dar una segunda oportunidad* *Si se equivoca en una, dar una la opción de que elija una tercera división*	Resuelve correctamente al menos dos de tres divisiones elegidas hasta en dos intentos 2564=64 328/8=41 219/3=73 225/5=45 328/4=82	Sí (1)	No resuelve correctamente al menos dos de tres divisiones elegidas en dos intentos.	No (0)

		204/6=34			
7	***Resuelve el siguiente problema y contesta la pregunta: ¿Cuánto dinero le quedó?*** *Si responde de manera incorrecta, preguntar si su resultado es definitivo. Si responde que no está seguro(a) y desea volver a hacer el problema, permitir realizarlo de nuevo.*	Proporciona la respuesta correcta: $32, hasta en dos intentos	Sí (1)	No proporciona la respuesta correcta hasta dos intentos	No (0)
8	***Resuelve el siguiente problema y contesta la pregunta ¿Cuánto gastó por todos los dulces que compró?*** *Si responde de manera incorrecta, preguntar si su resultado es definitivo. Si responde que no está seguro(a) y desea volver a hacer el problema, permitir realizarlo de nuevo.*	Proporciona la respuesta correcta: $315, hasta en dos intentos	Sí (1)	No proporciona la respuesta correcta hasta dos intentos	No (0)
9	***Elige dos fracciones y resuélvelas*** *Si se equivoca en una, dar una la opción de que elija una tercera fracción*	Resuelve correctamente al menos dos de tres fracciones elegidas hasta en dos intentos 3/8+5/7=61/56 ó 1 5/56 6/7+5/8=83/56 ó 1 27/56 2/5+3/6=27/30 1/4+8/9= 41/36 ó 1 5/36 5/6+3/7=53/42 ó 1 11/42 3/4+2/3=17/12 ó 1 5/12	Sí (1)	No resuelve correctamente al menos dos de tres fracciones elegidas en dos intentos	No (0)

Anexo 3. Análisis Factorial confirmatorio escala SMS asumiendo métrica continua

Cuadro 1. Análisis factorial confirmatorio del módulo SMS, escala completa

Jurisdicción A			
	Peso factorial	P-value	R2
math_sms1	0.475	0.000	0.226
math_sms2	0.649	0.000	0.421
math_sms3	0.676	0.000	0.457
math_sms4	0.659	0.000	0.434
math_sms5	0.495	0.000	0.245
math_sms6	0.606	0.000	0.367
math_sms7	0.262	0.000	0.069
math_sms10	0.428	0.000	0.183

RMSEA	0.064
CFI	0.945
TLI	0.923
SRMR	0.034

Jurisdicción B			
	Peso factorial	P-value	R2
math_sms1	0.458	0.000	0.209
math_sms2	0.578	0.000	0.334
math_sms3	0.640	0.000	0.410
math_sms4	0.699	0.000	0.488
math_sms5	0.553	0.000	0.306
math_sms6	0.602	0.000	0.362
math_sms7	0.306	0.000	0.093
math_sms10	0.447	0.000	0.200

RMSEA	0.056
CFI	0.957
TLI	0.939
SRMR	0.033

Jurisdicción C			
	Peso factorial	P-value	R2
math_sms1	0.567	0.000	0.322
math_sms2	0.529	0.000	0.280
math_sms3	0.650	0.000	0.422
math_sms4	0.653	0.000	0.427
math_sms5	0.589	0.000	0.347
math_sms6	0.544	0.000	0.296
math_sms7	0.390	0.000	0.152
math_sms10	0.468	0.000	0.219

RMSEA	0.086
CFI	0.907
TLI	0.869
SRMR	0.050

Fuente: Elaboración propia a partir de las bases de datos de línea de base 2021.

Cuadro 2. Análisis factorial confirmatorio del módulo SMS, escala reducida

Jurisdicción A			
	Peso factorial	P-value	R2
math_sms1	0.475	0.000	0.226
math_sms2	0.619	0.000	0.384
math_sms3	0.686	0.000	0.470
math_sms4	0.681	0.000	0.463
math_sms5	0.505	0.000	0.255

RMSEA	0.109
CFI	0.938
TLI	0.877
SRMR	0.041

Jurisdicción B			
	Peso factorial	P-value	R2
math_sms1	0.447	0.000	0.200
math_sms2	0.552	0.000	0.305
math_sms3	0.653	0.000	0.426
math_sms4	0.716	0.000	0.513
math_sms5	0.551	0.000	0.304

RMSEA	0.059
CFI	0.980
TLI	0.960
SRMR	0.024

Jurisdicción C			
	Peso factorial	P-value	R2
math_sms1	0.560	0.000	0.314
math_sms2	0.517	0.000	0.268
math_sms3	0.683	0.000	0.467
math_sms4	0.646	0.000	0.417
math_sms5	0.575	0.000	0.331

RMSEA	0.071
CFI	0.973
TLI	0.945
SRMR	0.027

Fuente: Elaboración propia a partir de las bases de datos de línea de base 2021.