ARISTAS 2022

REPORTE 17

ANÁLISIS DE LA MAGNITUD E INCIDENCIA DE LAS RESPUESTAS CON BAJO ESFUERZO ENTRE ESTUDIANTES DE TERCERO DE MEDIA

INEEd
Instituto Nacional de Evaluación Educativa

Aristas
Evaluación Nacional de Logros Educativos

La elaboración de este documento estuvo a cargo de: Betania Ávalos y Meliza González.

Corrección de estilo: Mercedes Pérez y Federico Bentancor
Diseño y diagramación: Diego Porcelli
Foto de tapa: DGES

Montevideo, 2025
ISSN: 2697-2786

Edificio Los Naranjos, Planta Alta, Parque de Innovación del LATU
Av. Italia 6201, Montevideo, Uruguay
(+598) 2604 4649 – 2604 8590
ineed@ineed.edu.uy
www.ineed.edu.uy

Cómo citar: INEEd. (2025). *Reporte de Aristas 17. Análisis de la magnitud e incidencia de las respuestas con bajo esfuerzo entre estudiantes de tercero de media*. Recuperado de https://www.ineed.edu.uy/images/Aristas/Publicaciones/Reportes/Reporte-17-Analisis-respuestas-con-bajo-esfuerzo-estudiantes-tercero-media.pdf

Este informe trata de adolescentes y adultos mujeres y varones. El uso del masculino genérico obedece a un criterio de economía de lenguaje y procura una lectura más fluida, sin ninguna connotación discriminatoria.

RESUMEN

Uno de los supuestos en los que se basan las evaluaciones educativas estandarizadas es que los estudiantes contestan realizando su máximo esfuerzo. Sin embargo, existe numerosa evidencia de que este supuesto no se cumple del todo, especialmente en pruebas de bajo impacto, como es el caso de Aristas.

Investigaciones han señalado que en el contexto de evaluaciones que presentan escasa o ninguna consecuencia personal, algunos estudiantes pueden carecer de motivación para hacer su mejor esfuerzo (Wise y Ma, 2012). Esta conducta introduce un sesgo negativo en el rendimiento de la prueba que puede disminuir la validez de las inferencias que se realizan sobre la base de los puntajes obtenidos (Wise y Gao, 2017)[1].

Por otro lado, existe evidencia de que el bajo esfuerzo o la baja motivación no afectan del mismo modo a todos los estudiantes, sino que ocurre diferencialmente de acuerdo a ciertas características como el género (Soland, 2018) y el riesgo de desvinculación (Soland y Kuhfeldb, 2019).

Como parte del monitoreo y revisión permanente de los procedimientos que preservan la rigurosidad de Aristas, el presente reporte analiza el comportamiento de bajo esfuerzo en dicha evaluación. En tal sentido, se plantea como objetivos: 1) identificar las respuestas con bajo esfuerzo en Aristas, 2) dimensionar su incidencia y 3) explorar la relación entre el bajo esfuerzo y las características sociodemográficas y socioemocionales de los estudiantes.

Se utilizaron los datos de respuestas a los ítems de las pruebas de matemática y lectura de la última aplicación de la evaluación Aristas en tercero de media[2], realizada en 2022. Se identificaron aquellas respuestas realizadas en un tiempo comparativamente muy breve, consideradas como indicativas de bajo esfuerzo, compatibles con lo que se ha denominado "respuesta de adivinación rápida".

Se calcularon medidas de esfuerzo con diferentes metodologías que presentaron resultados coherentes entre sí. En promedio, los estudiantes realizaron un 1,2% de las respuestas a los ítems de lectura y un 2,4% de matemática en un tiempo menor a 5 segundos. Los porcentajes de respuestas identificadas como adivinación rápida varían de acuerdo al método y el criterio utilizado, aunque en todos los casos fueron mayores en matemática que en lectura.

[1] Las pruebas de bajo impacto o bajo riesgo refieren a evaluaciones cuyos resultados tienen un impacto limitado o no tienen consecuencias directas para los estudiantes, los docentes ni los centros educativos en los que se lleva a cabo la evaluación. Las evaluaciones internacionales PISA y TERCE, así como la evaluación nacional Aristas son ejemplos de pruebas de este tipo.
[2] Con la transformación curricular integral ese grado pasó a llamarse noveno a partir de 2023.

Se observó que en ambas áreas la extraedad y el género masculino presentaron una relación negativa con el esfuerzo. Por su parte, el nivel socioeconómico del centro presentó una asociación positiva. En cuanto a las habilidades socioemocionales, se observó que la autorregulación metacognitiva y la perseverancia académica se asocian positivamente con el esfuerzo en ambas áreas. Al contrario, las habilidades de relacionamiento y las conductas externalizantes influyen de forma negativa.

Por último, el estudio discute sobre la importancia de continuar monitoreando el bajo esfuerzo, de manera de incorporar esta información para la mejor compresión de los desempeños de los estudiantes, así como mantener y reforzar medidas preventivas para reducir sus consecuencias.

ANTECEDENTES

JUSTIFICACIÓN

Uno de los supuestos en los que se basan las evaluaciones educativas es que los estudiantes responden realizando su máximo esfuerzo[3]. Sin embargo, existe evidencia de que esto no siempre es así, sino que la motivación y el esfuerzo varían, entre otros factores, por las consecuencias asociadas a la evaluación (Buchholz, Cignetti y Piacentini, 2022).

Investigaciones han señalado que en el contexto de evaluaciones que presentan escasa o ninguna consecuencia personal, algunos estudiantes pueden carecer de motivación para hacer su mejor esfuerzo (Wise y Ma, 2012). En efecto, este fenómeno ha sido analizado tanto en evaluaciones nacionales como internacionales. Respecto de las primeras, Valdivia et al. encontraron entre los estudiantes de octavo de un estado de Estados Unidos hasta un 8% que manifestaba bajo compromiso en sus respuestas a una evaluación interestatal (Valdivia, Rutkowski, Rutkowski, Canbolat y Underhill, 2023). Asimismo, Soland y Kuhfeldb (2019) estudiaron los resultados de una evaluación adaptativa de gran escala a lo largo de 4 años en un distrito del oeste de Estados Unidos, encontrando en promedio un 7,8% de estudiantes de tercero que presentaba respuestas caracterizadas como de adivinación rápida en matemática y un 17,7% en lectura.

En lo que respecta a evaluaciones internacionales, Ríos y Soland (2022) analizaron la respuesta de adivinación rápida en el dominio ciencias de la prueba PISA 2018, identificando un 6% de respuestas realizadas con este comportamiento entre todos los examinandos y aproximadamente un 6,3% para Uruguay. Por otro lado, utilizando una metodología diferente, un estudio de la Organización para la Cooperación y el Desarrollo Económicos (OCDE) (Avvisati, Buchholz, Piacentini y Vargas-Madriz, 2024) investigó acerca de la respuesta de adivinación rápida en el dominio de lectura de la prueba PISA 2018, encontrando que a nivel global un 7,4% de los estudiantes respondían en un tiempo inferior al considerado como mínimo, siendo esta cifra un 3,7% entre los adolescentes uruguayos.

Asimismo, el bajo esfuerzo o la baja motivación no afectan del mismo modo a todos los estudiantes. Existe evidencia de que ocurre diferencialmente por género (Soland, 2018) y a menudo se presenta más entre quienes abandonan los estudios (Soland y Kuhfeldb, 2019).

Además del estudio de la incidencia de la baja motivación, también se ha indagado en la relación de este tipo de conducta con características propias de los ítems, hallándose, por

[3] Entre los estándares que siguen las pruebas educativas de evaluación de programas o rendición de cuentas, se encuentra el hecho de que los examinandos realicen la prueba seriamente, con motivación y maximizando su esfuerzo (AERA, APA y NCME, 2018).

ejemplo, que la longitud del ítem es uno de los principales predictores del esfuerzo (Avvisati et al., 2024).

En cuanto a las consecuencias, el bajo esfuerzo introduce un sesgo negativo en el rendimiento de la prueba, lo que puede disminuir la validez de las inferencias que se realizan sobre la base de los puntajes obtenidos: menores niveles de esfuerzo o compromiso se asocian a desempeños más bajos (Wise y Gao, 2017). Al respecto, Wise y DeMars encontraron, a partir de una síntesis de resultados de varios estudios, que las puntuaciones de respondientes con baja motivación promediaban 0,58 desviaciones estándar más abajo que las de sus pares motivados (Wise y DeMars 2005, citado en Wise y DeMars, 2006).

De los distintos estudios realizados respecto de la temática se puede concluir que los resultados de las evaluaciones de bajo impacto representan una combinación de habilidades cognitivas y factores motivacionales (Buchholz et al., 2022). Estos hallazgos ponen de relevancia el análisis de las respuestas con bajo esfuerzo en las evaluaciones educativas de bajo impacto en general y, en nuestro caso, en el contexto de la aplicación de Aristas.

ACERCA DE LA MEDICIÓN DEL ESFUERZO EN EVALUACIONES EDUCATIVAS

El problema descrito ha sido desarrollado principalmente en publicaciones escritas en inglés y ha sido tratado de diferentes maneras, incluyendo: motivación (*test taker motivation*), esfuerzo (*test taker effort*) e indiferencia de los respondientes (*test taker disengagement*) (Avvisati et al., 2024). Si bien no se trata de un tema nuevo, su estudio se ha acrecentado a partir del paso de las evaluaciones en lápiz y papel a pruebas computarizadas[4], que han permitido desarrollar nuevos indicadores para su medición.

La investigación inicial respecto a esta temática se realizó principalmente a través de reportes autoadministrados al finalizar la evaluación, donde los mismos respondientes indicaban su nivel de compromiso o motivación durante la prueba (Wise y Gao, 2017). Como señalan Wise y Gao, este abordaje tiene al menos dos limitaciones. En primer lugar, no es posible saber qué tan veraces son las respuestas, dado que los evaluados podrían subestimar o sobrestimar el esfuerzo puesto en la prueba dependiendo de los efectos que puedan deducir de esta respuesta o también con base en su propio desempeño. En segundo lugar, su respuesta refiere a una visión general de la evaluación, que puede pasar por alto el rendimiento diferencial realizado en algunos momentos de la prueba.

Con la extensión de las pruebas por computadora se ha obtenido nueva información acerca de los procesos de respuesta a la evaluación, inimaginable en las realizadas con papel y lápiz. Esto ha permitido el desarrollo de otras formas de medir el esfuerzo, entre las cuales una de las más estudiadas ha sido el tiempo de respuesta al ítem (Wise y Gao, 2017).

[4] Metodología abreviada como CBA, por las iniciales de su denominación en inglés: *computer-based assessment*.

Se ha hallado evidencia de que, en el marco de pruebas de bajo impacto, el tiempo de respuesta es indicativo de la motivación de los examinandos. En particular, que las respuestas realizadas en un tiempo muy breve podrían ser "respuestas de adivinación rápida" (por su denominación en inglés: *rapid-guessing)*, asociadas a un bajo nivel de esfuerzo:

> El comportamiento de adivinación rápida puede ser identificado en aquellas respuestas que se realizan tan rápidamente que los examinados no tienen tiempo de considerar completamente el ítem. De este modo, las respuestas dadas en un contexto de adivinación rápida son esencialmente aleatorias y la probabilidad de que estén correctas será cercana al azar (Wise y Kong, 2005, p. 167)[5].

MÉTODOS DE IDENTIFICACIÓN

Umbral común

Las aproximaciones a medir el esfuerzo de los respondientes a partir de los tiempos de respuesta son variadas. Entre las más simples se encuentran aquellas que establecen un umbral mínimo de tiempo de respuesta común a todos los ítems, bajo el cual se considera que se trata de una respuesta de adivinación rápida. A modo de ejemplo, en el análisis de las pruebas PISA, la OCDE ha llevado a cabo estudios utilizando un umbral común de 5 segundos por ítem (Avvisati et al., 2024).

Umbral normativo

Otro enfoque bastante utilizado es el método del "umbral normativo" desarrollado por Wise y Ma (2012). Este establece un umbral mínimo de tiempo de respuesta para cada ítem, calculado como un porcentaje del tiempo medio de respuesta a este. De esta manera, lo que puede considerarse respuesta de adivinación rápida varía dependiendo de la demanda de tiempo propia de cada ítem. A su vez, considera un tiempo máximo posible[6]. Las respuestas realizadas en un tiempo inferior al establecido por el umbral son marcadas como respuestas de adivinación rápida. La abreviación NT10 (de *normative threshold*, umbral normativo) se utiliza para indicar el umbral que considera el 10% del tiempo promedio de respuesta, que constituye uno de los umbrales más usados. Otros valores de referencia, menos conservadores[7], son por ejemplo el NT20 (20%) y el NT30 (30%). A partir de estas medidas se puede calcular un índice general de esfuerzo en el tiempo de respuesta para cada estudiante, el cual corresponde a la proporción de ítems de la prueba en que el tiempo de respuesta fue superior al valor establecido por el umbral (Guo et al., 2016).

---

[5] La traducción es nuestra. El texto original es: "Rapid guessing behavior can be identified by responses occurring so rapidly that examinees did not have time to fully consider the item. Thus, answers given during rapid-guessing behavior are essentially random, and the correctness of these answers will be at or near chance levels".

[6] En esta publicación establecen como máximo 10 segundos para los umbrales. En otras aplicaciones se han realizado variaciones de este criterio, por ejemplo, el estudio de Ríos y Soland (2022) no considera valores máximos.

[7] La mayor restricción en el umbral mínimo de tiempo de respuesta por parte del NT10 en comparación al NT20 o al NT30 lo sitúa como un criterio más conservador, en el sentido de que clasifica como respuestas de adivinación rápida aquellas que inequívocamente se identifican como tales, sin embargo, clasifica como respuestas con esfuerzo aquellas que toman solo un poco más de tiempo, más ambiguas en cuanto a su interpretación. Al contrario, el NT20 y el NT30 amplían el umbral de tiempo para considerar una respuesta como de adivinación rápida.

Mezcla de log-normales

Una tercera aproximación es la definición del umbral mínimo a partir de una inspección visual de la distribución de los tiempos de respuesta (Wise y DeMars, 2006). Se ha estudiado que, en presencia de respuestas de bajo esfuerzo, los tiempos de respuesta presentan una distribución bimodal: se observa un primer "pico" situado en un tiempo de respuesta relativamente bajo, seguido de una distribución que agrupa mayor frecuencia de respuestas y que presenta una moda de tiempo mayor. El umbral para diferenciar las respuestas de adivinación rápida se define identificando el punto mínimo entre las dos distribuciones. Una adaptación de este método es la basada en la mezcla de distribuciones log-normales (denominado MLN). El método MLN emplea un proceso automatizado que utiliza la distribución empírica del tiempo de respuesta, asumiendo que presenta una distribución log normal mixta y, a continuación, localiza el punto que minimiza la mezcla (intersección de las densidades), que se establece como umbral (Rios y Guo, 2020).

Proporción acumulada

Guo et al. (2016) plantean un nuevo método (abreviado como CUMP) que no solo considera los tiempos de respuesta, sino también el nivel de acierto de las respuestas. En esta aproximación las respuestas de bajo esfuerzo son aquellas cuya probabilidad de ser correctas se encuentra por debajo de la elección al azar. Así, para cada ítem se calcula la proporción acumulada de respuestas correctas considerando los tiempos de respuesta. La proporción de respuestas correctas acumulada al tiempo t tiene una tendencia estable que converge a la proporción global de respuestas correctas (dificultad del ítem) a medida que aumenta t. De esta manera, es posible establecer un umbral de tiempo de respuesta en el momento exacto en que la proporción acumulada supera el acierto propio del azar (por ejemplo, 0,25 en el caso de un ítem de respuesta múltiple con cuatro opciones de respuesta). Una limitación de este método es que no aplica para ítems que resulten muy difíciles y presenten una proporción total de respuestas correctas inferior o similar a la del acierto al azar (Rios y Guo, 2020).

Cuantiles de tiempo de respuesta a los ítems

Una aproximación particular es la utilizada por Guo et al. (2016), denominada RT Quantile Analysis, que genera una medida resumen para toda la prueba a partir de los tiempos de respuestas de todos los ítem. En concreto, se identifica visualmente un umbral general a partir de la distribución de los cuantiles de tiempos de respuesta de los ítems, cuando se presenta una distribución bimodal. De esta manera, se distinguen como de bajo esfuerzo a las aplicaciones con una mediana de tiempo de respuesta por debajo del umbral.

Tal como señalan Papanastasiou y Michaelides, las comparaciones acerca de un enfoque óptimo de identificación de umbrales no han sido concluyentes y no hay consenso respecto a un método preferido, ya que cada uno presenta fortalezas y debilidades (2024).

ARISTAS

La evaluación nacional de logros educativos Aristas se realizó por primera vez en tercero y sexto de primaria en 2017 y en tercero de media en 2018. La prueba produce información sobre los desempeños de los estudiantes en lectura y matemática, con foco en los conocimientos y capacidades que el sistema educativo uruguayo se propone que alcancen. Además, busca conocer las habilidades socioemocionales de los estudiantes y sus opiniones sobre el clima escolar, la convivencia y la participación en los centros. Al mismo tiempo, pregunta a directores y docentes acerca de las formas habituales de dar clase y recoge sus puntos de vista sobre el clima escolar y el trabajo en los centros educativos[8].

Con el objetivo de constituirse como un instrumento de seguimiento continuo, se realiza cada tres años en ambos subsistemas. Gracias a la infraestructura tecnológica de los centros educativos de Uruguay debido al trabajo de Ceibal, a partir del primer año de educación primaria cada alumno cuenta con una computadora portátil propia. Esto facilitó que desde la primera aplicación las pruebas Aristas se realizaran en computadora.

[8] Adicionalmente, para el caso de los alumnos de primaria se consulta a sus familias para conocer el contexto familiar de estos niños.

OBJETIVOS

Desde sus inicios, el INEEd ha trabajado en la mejora y la actualización de los distintos procesos involucrados en la aplicación de Aristas, cuidando al mismo tiempo de mantener la comparabilidad. Como parte del monitoreo y la revisión permanente de los procedimientos que preservan la rigurosidad de esta evaluación, este estudio se propone:

- identificar las respuestas de adivinación rápida en la evaluación Aristas Media, analizando diferentes métodos de aproximación;

- dimensionar la incidencia del bajo esfuerzo entre los estudiantes participantes de Aristas Media, y

- evaluar el vínculo del esfuerzo en realizar la prueba con las características sociodemográficas y habilidades socioemocionales de los estudiantes.

La aproximación a la identificación de comportamientos con bajo esfuerzo se realizará a partir de los tiempos de respuesta de los estudiantes. Al respecto, el desafío es distinguir tantas respuestas de adivinación rápida como sea posible, sin identificar erróneamente las que no lo son, de manera de estimar este tipo de comportamiento y sus características.

Si bien la respuesta de *adivinación rápida* mide una faceta específica del comportamiento de *bajo esfuerzo,* más amplio, en el marco de este informe utilizaremos ambos términos indistintamente, ya que la adivinación rápida es un indicador de bajo esfuerzo.

DATOS

Se utiliza la información de las pruebas de lectura y matemática de la última aplicación definitiva de la evaluación Aristas Media, llevada adelante para tercero en 2022. Con respecto a sus condiciones de realización, es importante señalar que estas pruebas son administradas por computadora y se efectúan en su gran mayoría utilizando las máquinas de los estudiantes o el centro educativo[9]. Se llevan adelante bajo la instrucción y supervisión de aplicadores externos[10].

El tiempo máximo establecido para responder la prueba es de 70 minutos, aunque se realizan en un promedio aproximado de 30 minutos. La distancia entre el tiempo máximo y el de realización efectiva permite que la prueba sea realizada al ritmo de cada estudiante, cobrando pertinencia la utilización de los tiempos de respuesta para dar cuenta de la adivinación rápida, ya que no existe una presión en lo que respecta a los tiempos de respuesta.

Previo al inicio de la evaluación, los estudiantes ingresan a la plataforma informática en que realizarán la prueba y tienen un breve ejercicio didáctico para familiarizarse con la plataforma y el formato de los ítems múltiple opción. Una vez comenzada la prueba no es obligatorio dar una respuesta para continuar avanzando, se pueden dejar ítems sin resolver y contestarlos después. Antes de finalizar, la plataforma muestra un resumen de las respuestas donde se indican las preguntas sin responder, a las cuales se puede acceder directamente.

La aplicación completa de Aristas en un centro educativo lleva dos días, uno para cada prueba. En el caso de efectuarse además algún cuestionario, este se realiza después de la prueba, mediando un momento de recreo.

Con respecto a los tiempos de respuesta, este estudio utiliza la información disponible, que actualmente es el registro de la última interacción del evaluado con el ítem[11].

---

[9] En caso de que el centro educativo no cuente con computadoras suficientes para asegurar una por estudiante, el equipo de campo del INEEd coordina previamente el envío de dispositivos de préstamo para realizar la aplicación.
[10] Los aplicadores externos son contratados por el INEEd. Pueden resolver consultas acerca de los detalles operativos de la aplicación, sin embargo, no cuentan con información ni pueden contestar preguntas acerca del contenido de las pruebas.
[11] Para las aplicaciones a partir de 2025 se contará con información sobre todo el recorrido por la prueba, lo que permitirá conocer los tiempos involucrados en todas las interacciones con los ítems, no solo la última. Los datos de la aplicación piloto de Aristas Media 2024 mostraron que el 94% de las respuestas fueron realizadas en una sola interacción con el ítem. Otra referencia al respecto es el estudio de Rios y Soland (2022), donde analizan la respuesta rápida en la prueba de ciencias de PISA 2018 únicamente a partir de la última interacción con el ítem, dado que sus análisis demostraron que los respondientes interactuaban solo una vez con el ítem en el 93% de los casos, y la mediana de tiempo de respuesta era prácticamente idéntica considerando la última interacción o el total del tiempo invertido en el ítem.

ÍTEMS

Si bien cada prueba difiere en la cantidad total de ítems, su estructura es similar: en cada evaluación se trabaja con un total de entre 150 a 280 ítems que se distribuyen en cerca de 15 cuadernillos, cada uno con entre 20 y 30 ítems. Existen tres formatos de ítems: cerrados de múltiple opción (CE), interactivos de correlación (CO) y de respuesta construida (AD), siendo bastante mayoritario el tipo cerrado múltiple opción. La tabla 1 presenta la distribución de los ítems de cada prueba según su tipo y en los cuadernillos.

TABLA 1
DISTRIBUCIÓN DE LOS ÍTEMS SEGÚN TIPO, POR ÁREA Y GRADO

Grado y área		Ítems				Cuadernillos	
		N total	CE	AD	CO	N	Ítems
Tercero	Lectura	280	260	20	0	20	28: 26 CE, 2 AD
	Matemática	240	224	16	0	16	30: 28 CE y 2 AD

Fuente: elaboración propia a partir de datos de Aristas Media 2022.

Debido a la presencia ampliamente mayoritaria de los ítems cerrados de múltiple opción en todas las pruebas, así como a la mayor aplicación de metodologías asociadas al tiempo de respuesta en este tipo de ítems, se optó por aplicar este estudio de identificación de respuestas de adivinación rápida únicamente a este formato de ítems[12].

DEPURACIÓN

En todos los análisis se depuraron en primer lugar los casos eliminados regularmente de la muestra de Aristas (incidencias de campo, duplicados y casos que no visualizaron ningún ítem) y los estudiantes con necesidades educativas específicas. Luego, para este estudio en particular se eliminaron, además, los siguientes: 1) los estudiantes que no vieron todos los ítems de la prueba (0,2% de las aplicaciones) y 2) los estudiantes que respondieron los ítems cerrados de múltiple opción en un tiempo total superior a una hora (3.600 segundos) (0,8% de las aplicaciones).

[12] Al respecto, cabe señalar la investigación de Wise y Gao (2017), donde se encontró que la respuesta rápida a ítems de múltiple opción constituyó el principal comportamiento de bajo esfuerzo en un estudio que consideró a su vez la omisión rápida a este mismo tipo de ítems y la respuesta rápida-superficial (*rapid perfunctory answering*) a ítems de respuesta construida.

ANÁLISIS

En esta sección se presenta el análisis de los resultados, detallando las consideraciones metodológicas. Se comienza con la descripción de los tiempos de respuesta a los ítems de cada prueba. A continuación, se especifican los métodos empleados para identificar las respuestas de adivinación rápida y se analiza su desempeño en esta muestra. Luego, se estiman porcentajes de este tipo de comportamiento detectados en los participantes de Aristas Media. Posteriormente, se exploran las características de los estudiantes que se relacionan con los índices de esfuerzo.

DESCRIPCIÓN DE LOS TIEMPOS DE RESPUESTA TOTALES Y POR ÍTEM

A continuación, se presenta un resumen (tabla 2) que describe los tiempos totales de la prueba y los referidos a los ítems de múltiple opción, para cada área. Como puede observarse, la duración total de una prueba promedia 30 y 34 minutos (1.805 y 2.071 segundos) para lectura y matemática, respectivamente. Si solo se considera el tiempo asociado a los ítems de múltiple opción, el promedio es de 25 minutos en lectura (1.488 segundos) y 30 en matemática (1.775 segundos). Con respecto a los ítems, se observa un promedio de cerca de 2 minutos por ítem, y una mediana de alrededor de 50 segundos.

TABLA 2
DISTRIBUCIÓN DEL TIEMPO DE RESPUESTA DE LOS ESTUDIANTES A LAS PRUEBAS, MEDIDO EN SEGUNDOS, POR ÁREA

Grado y área		Muestra	Tempo total prueba		Tiempo total ítems CE		Media del tiempo ítems CE		Mediana del tiempo ítems CE	
			Media	Desvío	Media	Desvío	Media	Desvío	Media	Desvío
Tercero	Lectura	9.530	1.805	634	1.488	536	110	40	42	18
	Matemática	9.358	2.071	744	1.775	656	122	45	53	21

Fuente: elaboración propia a partir de datos de Aristas Media 2022.
Nota: el tiempo total de prueba se calcula a partir de las horas de inicio y finalización de la aplicación del estudiante.

IDENTIFICACIÓN DE BAJO ESFUERZO

MÉTODOS UTILIZADOS

Para detectar respuestas de adivinación rápida o bajo esfuerzo, se emplearon varios de los métodos referenciados y se comparó su desempeño.

1. En primer lugar, se identificó visualmente un umbral genérico por prueba a partir del análisis de los cuantiles del tiempo de respuesta de los estudiantes al total de ítems CE (**RT cuantil 50**). En este caso, el resultado es una variable dicotómica a nivel de prueba, que toma valor 1 cuando la mediana de tiempo promedio de respuesta de un estudiante a todos los ítems está por debajo del umbral y 0 cuando lo supera.

Adicionalmente, se emplearon métodos para identificar adivinación rápida en la respuesta a cada ítem. Los criterios explorados fueron los siguientes:

2. Un **umbral común** de 5 segundos para todos los ítems[13].

3. También se aplicó un método basado en mezcla de distribuciones log-normales donde se utilizó el algoritmo esperanza-maximización EM para calcular automáticamente el umbral entre las modas de cada ítem. Este último enfoque se complementó con la técnica de proporción acumulada (estableciendo el valor 0,25 como proporción de acierto al azar) que se usó en aquellos ítems donde el umbral calculado no resultaba adecuado. A raíz del uso conjunto de estos dos métodos, esta técnica se llamó **método híbrido**, similar a como la refieren Rios y Guo (2020).

4. Además, se exploró el método del **umbral normativo**, empleando umbrales del 10%, 15% y 20% del tiempo medio de cada ítem (NT10, NT15 y NT20, respectivamente).

En todos los casos, se estableció un mínimo de 3 segundos y un máximo de 20 segundos[14] para los umbrales hallados.

A partir de estos métodos se determinó un umbral de tiempo por ítem que permitió identificar aquellas respuestas que se realizaron en un tiempo inferior al umbral establecido, las cuales se consideraron como adivinación rápida. Posteriormente, se crearon índices de esfuerzo de los estudiantes según cada método (RTE[15]), calculados como el promedio de respuestas no marcadas como adivinación rápida:

$$RTE = \frac{\text{número de ítems con tiempo de respuesta sobre el umbral}}{\text{número total ítems}},$$

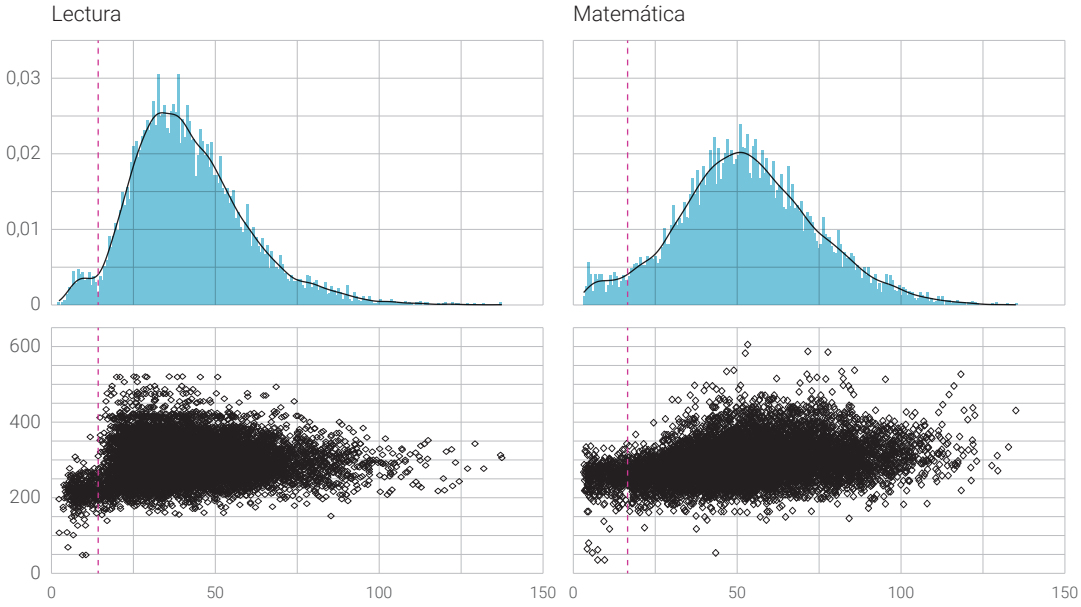
donde el umbral se define para cada ítem según el método utilizado.

Se evalúan dos valores del índice RTE para clasificar a estudiantes que realizan la prueba con bajo esfuerzo: (1) valores inferiores a 0,9 (estudiantes con más del 10% de los ítems con respuesta de adivinación rápida) y (2) valores inferiores a 0,7 (más del 30% de los ítems de adivinación rápida)[16].

[13] La elección de este umbral toma como referencia los estudios realizados por la OCDE con los datos de PISA, donde utilizan el límite de 5 segundos (Avvisati et al., 2024).
[14] En los estudios revisados las metodologías contemplaban el uso de mínimos y máximos. El mínimo de 3 segundos fue escogido en el entendido de que un tiempo inferior no permite interactuar mínimamente con el ítem; por su parte, el máximo de 20 segundos fue seleccionado, entre otros umbrales menores, por permitir mayor variabilidad entre ítems.
[15] Sigla de *response time effort*, es la abreviación que se le da a este índice en la literatura especializada.
[16] Estos valores fueron elegidos tomando en cuenta que los estudios en la temática señalan que valores 0,9 e inferiores indican puntajes sesgados que cuestionan la validez de las estimaciones (Wise, 2015), mientras que 0,7 es un criterio más conservador utilizado para depurar casos (Soland, 2018; Wise y Kong, 2005).

El gráfico 1 presenta la aplicación del primer método (RT cuantil 50). El panel superior muestra el histograma de la mediana del tiempo de respuesta por ítem en las pruebas de tercero, mientras que el panel inferior presenta esta información cruzada con la habilidad estimada del estudiante. La línea vertical punteada indica el umbral RT (14 segundos para lectura y 16 para matemática), por debajo del cual se considera que la prueba fue realizada con bajo esfuerzo. Como puede observarse, en el caso de lectura este umbral coincide con un claro incremento de la variabilidad de la habilidad, ilustrado en el panel inferior por valores más altos que alcanza la nube a la derecha del umbral. Sin embargo, en matemática el límite de la nube de puntos en torno del umbral es más difuso[17].

GRÁFICO 1
HISTOGRAMA DE LA MEDIANA DEL TIEMPO DE RESPUESTA POR ÍTEM (PANEL SUPERIOR). MEDIANA DEL TIEMPO DE RESPUESTA A LOS ÍTEMS DE LA PRUEBA SEGÚN PUNTAJE DE HABILIDAD (PANEL INFERIOR)
AÑO 2022
Informantes: estudiantes de tercero de media



Fuente: elaboración propia a partir de datos de Aristas Media 2022.
Nota: los puntajes de las pruebas se expresan en una escala de media 300 y desvío 50.

Como resultado de los métodos umbral normativo e híbrido, se obtuvieron umbrales por ítem. En el gráfico A.1 del Anexo se presenta el histograma de los umbrales según método para cada una de las pruebas. De allí puede destacarse una alta proporción de ítems con umbrales igual al límite 20 en el método híbrido, en particular en matemática, en su mayoría aportados por el enfoque de proporción acumulada, siendo un indicador de no ajuste del método, en tanto no aporta umbrales diferenciados.

La poca adecuación del enfoque de proporción acumulada asignando umbrales máximos en el caso de la prueba de matemática se asocia al nivel general de acierto relativamente bajo (0,38) en esta área o al menos más cercano al nivel que este método establece como propio del azar (0,25). En este caso existe el riesgo de que se categoricen como respuestas de

[17] En general, puede observarse que los estudiantes identificados con dicho criterio presentan en su gran mayoría una habilidad estimada menor a la media 300.

bajo esfuerzo simplemente a tentativas de respuesta erradas, pero que no necesariamente fueron hechas con bajo esfuerzo.

La tabla 3 resume los resultados de todos los métodos. Allí se presenta el tiempo promedio dedicado a los ítems y los umbrales hallados por las diferentes metodologías. Los umbrales por prueba seleccionados visualmente a partir del cuarto método (RT cuantil 50) fueron 14 y 17 segundos para la mediana del tiempo de respuesta a los ítems de lectura y matemática, respectivamente. Por su parte, se observa que el valor promedio de los umbrales calculados por ítem se encuentra en el rango entre 5,8 y 12,5 segundos según la prueba y el criterio utilizados.

TABLA 3
MEDIA (Y DESVÍO) DE LOS TIEMPOS DE RESPUESTA EN SEGUNDOS PROMEDIO POR ÍTEM, MEDIA (Y DESVÍO) DE LOS UMBRALES DE LOS ÍTEMS OBTENIDOS POR LOS MÉTODOS NT E HÍBRIDO Y UMBRAL POR PRUEBA OBTENIDO POR MÉTODO RT CUANTIL 50, SEGÚN ÁREA

Área	Tiempo medio por ítem	Umbral por prueba	Media y desvío umbrales por ítem			
		RT cuantil 50	NT10	NT15	NT20	Híbrido
Lectura	57,2(35,2)	14	5,8(3,5)	8,3(4,2)	10,5(4,4)	8,2(4,5)
Matemática	63,4(18,7)	17	6,3(1,8)	9,5(2,8)	12,5(3,5)	11,1(6)

Fuente: elaboración propia a partir de datos de Aristas Media 2022.

En el Anexo se detallan las pruebas de validación realizadas a las medidas de esfuerzo (RTE) calculadas según los métodos de identificación mencionados, así como la clasificación obtenida por el método RT cuantil. A partir de estas pruebas es posible obtener las siguientes conclusiones:

- Los resultados cumplen criterios de validez como medida de esfuerzo e identificación de respuestas de adivinación rápida, dado que se corroboran los siguientes criterios:
    - los índices RTE se correlacionan con la habilidad del estudiante y el tiempo dedicado a la prueba;
    - los estudiantes clasificados con bajo esfuerzo obtienen aciertos similares a los que se esperarían al azar y significativamente inferiores a los de los estudiantes que realizaron la prueba con mayor esfuerzo, independientemente de su nivel socioeconómico, y
    - por último, si se analiza el efecto de depurar estos casos en la correlación de variables teóricamente asociadas, se espera que al eliminarlos la correlación aumente: en efecto, las pruebas de depuración según el índice de esfuerzo muestran un aumento en la correlación entre la habilidad y el nivel socioeconómico y cultural, lo que sugiere que los puntajes de habilidad de los casos clasificados con bajo esfuerzo podrían presentar una menor validez.
- Los índices obtenidos a partir de los diferentes métodos tienen comportamientos similares y consistentes.

- Los métodos más conservadores (umbral común y NT10) presentan menor sensibilidad en comparación con el resto (NT15, NT20, híbrido).

- La clasificación de bajo esfuerzo a partir del criterio RTE<0,7 se corresponde con resultados de adivinación rápida con respecto a la proporción de aciertos de la prueba cercana al azar, en mayor medida que el criterio RTE<0,9.

INCIDENCIA DE LAS RESPUESTAS CON BAJO ESFUERZO

La incidencia de respuestas con bajo esfuerzo en las pruebas de Aristas puede reportarse de múltiples formas y las estimaciones dependen del método y la exigencia del criterio. Por este motivo, la interpretación, así como la comparación con otros estudios, debe hacerse a la luz del método utilizado.

En esta sección se presentan varias estadísticas que permiten dimensionar la incidencia de este tipo de comportamiento, considerando cada uno de los métodos explorados. En la interpretación de dichos resultados se resaltará el método NT10 como método conservador para identificar respuestas de bajo esfuerzo, ya que considera de manera más precisa las características del ítem en comparación con el umbral común de 5 segundos. Sin embargo, es importante señalar que los índices obtenidos por las otras metodologías son válidos y pueden ser útiles para análisis específicos, siendo algunos relevantes para la comparación con otros estudios.

La tabla 4 presenta estimaciones para los estudiantes, ponderados por el peso del estudiante en la muestra. Primero se presenta el valor promedio de los índices de esfuerzo de los estudiantes (RTE), donde se observan en general promedios elevados (sobre 0,9). Según el método NT10, puede interpretarse que los estudiantes respondieron con bajo esfuerzo, en promedio, un 1,3% de los ítems de lectura y un 3,5% de los de matemática.

Adicionalmente, se presentan los porcentajes de estudiantes que tuvieron al menos una respuesta con adivinación rápida (RTE<1). De allí se puede señalar que, según los métodos más conservadores, más de un 90% en lectura y un 80% en matemática realizan el total de la prueba sin ninguna respuesta con adivinación rápida. Además, se presentan los porcentajes de estudiantes con 10% o más de las respuestas realizadas con adivinación rápida (RTE<0,9) y 30% o más (RTE<0,7). Estos porcentajes pueden interpretarse como diferentes aproximaciones a medir la incidencia del bajo esfuerzo en la realización de la prueba. Se observa que en el primer caso (RTE<0,9) el porcentaje de estudiantes con bajo esfuerzo asciende a 4,6% en lectura y 11,6% en matemática, usando el criterio NT10; en el segundo caso (RTE<0,7) el porcentaje es 0,7% en lectura y 3,3% en matemática, con el mismo método. Por su parte, según el método RT cuantil 50, que clasifica estudiantes según el esfuerzo en tiempo dedicado en general a la prueba, la incidencia en ambas áreas es similar (3,6% y 4,6%, respectivamente).

TABLA 4

PROMEDIO RTE Y PORCENTAJE DE ESTUDIANTES CON VALORES INFERIORES A 1, 0,9 Y 0,7 POR ÁREA, SEGÚN MÉTODO

	Área	Umbral común	NT10	NT15	NT20	Híbrido
RTE del estudiante en promedio	Lectura	0,988	0,987	0,969	0,952	0,964
	Matemática	0,976	0,965	0,941	0,922	0,928
Porcentaje de estudiantes con RTE<1	Lectura	9,7%	12,8%	20,6%	30,3%	27,1%
	Matemática	18,2%	24,9%	34,7%	43,7%	51,9%
Porcentaje de estudiantes con RTE<0,9	Lectura	4,2%	4,6%	9,4%	13,3%	10,8%
	Matemática	7,7%	11,6%	18,1%	22,9%	22,9%
Porcentaje de estudiantes con RTE<0,7	Lectura	1,1%	0,7%	3,4%	5,3%	3,7%
	Matemática	2,0%	3,3%	6,2%	8,3%	6,7%
Porcentaje de estudiantes métodos RT cuantil 50	Lectura	3,6%				
	Matemática	4,6%				

Fuente: elaboración propia a partir de datos de Aristas Media 2022.
Nota 1: RTE es el índice de esfuerzo con base en el tiempo de respuesta, da cuenta de la proporción de ítems que cada estudiante realizó sin respuesta de adivinación rápida.
Nota 2: cantidad total de estudiantes: 9.530 en lectura y 9.358 en matemática.

A modo de síntesis, se observa que en todos los métodos la incidencia es mayor en matemática que en lectura. En términos generales, los resultados varían según el método y, como es de esperar, los más conservadores (umbral común y NT10) dan como resultado menor incidencia de adivinación rápida, mientras que las cifras crecen a medida que se consideran criterios que aumentan los umbrales (NT15, NT20). Asimismo, las estimaciones también dependen del porcentaje de respuestas de adivinación rápida que se considere indicativo de bajo esfuerzo en la prueba (RTE<1; 0,9; 0,7).

CARACTERIZACIÓN DE LOS ESTUDIANTES CON RESPUESTAS DE BAJO ESFUERZO

Para explorar la relación entre el bajo esfuerzo y diversas variables de caracterización de los estudiantes, se llevaron a cabo dos modelos multinivel. En estos modelos, el nivel 1 corresponde a los estudiantes y el 2 a los grupos a los que pertenecen.

Se estimó un modelo para cada área utilizando el índice de esfuerzo NT15 como variable dependiente y las siguientes variables explicativas:

- contexto socioeconómico del centro,
- estatus socioeconómico del estudiante,
- género,
- extraedad[18],
- región (Montevideo/interior) del centro educativo y

[18] Se considera con extraedad a todo aquel estudiante matriculado en un determinado período curricular en educación común que al 30 de abril del mismo período excede en al menos un año la edad teórica correspondiente al grado (tomado de las definiciones del Monitor Educativo).

- habilidades socioemocionales de los estudiantes (autorregulación metacognitiva, motivación intrínseca, perseverancia académica, autoeficacia académica, valoración de la tarea, empatía, habilidades de relacionamiento, regulación emocional, autocontrol, conductas internalizantes y conductas externalizantes)[19].

Todas las variables de los modelos se incluyen estandarizadas para facilitar la comparación.

El gráfico 2 muestra la influencia de las variables del modelo que presentan una asociación significativa con un 95% de confianza respecto del índice de esfuerzo, según área. En la tabla A.4 del Anexo se amplía información sobre el modelo.

GRÁFICO 2
RESULTADOS DEL MODELO DE REGRESIÓN LINEAL MULTINIVEL EN EL ÍNDICE DE ESFUERZO EN PRUEBAS DE LECTURA Y MATEMÁTICA

Fuente: elaboración propia a partir de datos de Aristas Media 2022.
Nota 1: las variables se encuentran estandarizadas.
Nota 2: la variable dependiente del modelo es RTE método NT15.
Nota 3: se grafican los coeficientes que resultan significativos al 95% ($p<0,05$).

[19] Estas variables corresponden a índices que dan cuenta de las subdimensiones de habilidades socioemocionales que se relevan en Aristas. En la tabla A.3 del Anexo se señala su definición y la dimensión a la que pertenecen.

Se destaca que la extraedad es la variable que se relaciona más negativamente con el índice de esfuerzo en todas las áreas. En el caso del género, se observa en las dos áreas que ser varón tiene una asociación negativa con el esfuerzo, principalmente en matemática. En lectura también muestran menor esfuerzo los estudiantes que declararon otro género[20].

Se observa que existe una asociación positiva entre el nivel socioeconómico del centro y el índice de esfuerzo. El estatus socioeconómico de la familia pierde relevancia al controlar por el contexto del centro, siendo no significativo en matemática y con un efecto pequeño, con signo negativo, en lectura. Por otra parte, la pertenencia a un centro educativo del interior presenta una asociación positiva con el índice de esfuerzo, solo en lectura.

En lo que respecta a los índices de habilidades socioemocionales, observamos que los índices de conductas internalizantes, autoeficacia académica, perseverancia académica y autorregulación metacognitiva presentan una asociación positiva con el esfuerzo. En el caso de la perseverancia académica, se destaca que esta relación únicamente es significativa en el caso de matemática, donde presenta una influencia mucho más fuerte que todas las otras variables consideradas en el modelo. Por su parte, tanto las habilidades de relacionamiento como las conductas externalizantes (que refieren a comportamientos asociados a la hiperactividad, agresividad y conductas oposicionistas) presentan una relación negativa con el esfuerzo, más en matemática que en lectura.

[20] En el caso de tercero de media, la consulta por género tiene como opciones de respuesta: "varón", "mujer" y "otro". El porcentaje de estudiantes que se autoidentifica con el género "otro" constituye el 2,3% del total.

SÍNTESIS Y DISCUSIÓN

Este trabajo constituye un primer estudio exploratorio acerca de las respuestas de bajo esfuerzo en Aristas Media. Con este objetivo se han investigado distintas técnicas de identificación, tomando como base los tiempos de respuesta.

A partir de los análisis realizados se destacan cuatro tópicos para la discusión.

1. **Aproximaciones a la identificación de respuestas de adivinación rápida**

Se exploraron métodos que determinan umbrales del tiempo de respuesta para cada ítem de la prueba:

- umbral común (5 segundos);
- umbral normativo NT10, NT15 y NT20 (basado en el 10, 15 y 20% del tiempo medio de respuestas a los ítems, respectivamente), y
- un método híbrido que combina el enfoque de mezcla de distribuciones log-normales y el de proporción acumulada.

A partir de estos se calcularon índices de esfuerzo basados en el tiempo de respuesta, según la proporción de ítems respondidos en tiempos superiores al umbral.

Además, se empleó un método que determina visualmente un umbral único por prueba, a partir de la distribución bimodal de la mediana del tiempo de respuesta a todos los ítems de la prueba (RT cuantil 50).

Los resultados en general fueron coherentes entre sí. Se emplean los puntos de corte genéricos 0,7 y 0,9 en los índices de esfuerzo para clasificar a los estudiantes de bajo esfuerzo. Se observa que los estudiantes con índices de esfuerzo menores a 0,7 (30% o más de ítems respondidos con adivinación rápida) tienen una proporción de aciertos similar a la respuesta al azar, independientemente de su nivel socioeconómico.

Al evaluar el desempeño general de los métodos en las diferentes áreas considerando los objetivos de este estudio, para hacer seguimiento del comportamiento se prefieren índices conservadores como el NT10, que mostró un comportamiento similar al umbral común de 5 segundos, pero considerando diferencias en los ítems. Los índices calculados a partir de métodos más exigentes (NT15, NT20, híbrido) mostraron validez y mayor sensibilidad para medir el esfuerzo, por lo que pueden ser útiles para otros análisis específicos que contribuyan a comprender las habilidades académicas de los estudiantes. El método RT

también resulta valioso por su simplicidad y coherencia en la clasificación de estudiantes de bajo esfuerzo.

2. **Incidencia de bajo esfuerzo**

Los análisis realizados muestran que los porcentajes de adivinación rápida obtenidos en Aristas Media están en general dentro de lo esperado en este tipo de evaluaciones.

Dependiendo del método utilizado, ya sea el más o menos conservador, y según el área evaluada, se estima que, en promedio, entre el 1,2% y el 8% de las respuestas de los estudiantes son producto de adivinación rápida. Estas estimaciones varían según la exigencia del criterio aplicado. Es importante señalar que, si bien este reporte prioriza ciertas metodologías para la interpretación de los resultados, otras también son relevantes por su comparabilidad con estudios previos.

De acuerdo a un método conservador como el NT10, considerando un RTE<0,7 (es decir, estudiantes que responden al menos 30% de los ítems de la prueba en un tiempo inferior al 10% del tiempo promedio del ítem), se observa que solo un 0,7% de los estudiantes realizó un bajo esfuerzo en lectura, mientras que en matemática la cifra asciende al 3,3%.

El informe de la OCDE reporta para los estudiantes uruguayos que participaron de PISA 2018 una proporción promedio de ítems respondidos empleando más de 5 segundos por ítem de 0,975 en matemática y ciencias[21] (analizando ambos dominios en conjunto) (OCDE, 2019). En Aristas Media 2022, con el mismo criterio (5 segundos), se obtuvieron promedios similares: 0,987 en lectura y 0,975 en matemática, es decir, en promedio los estudiantes adivinan un 1,3% de las respuestas en lectura y un 2,5% en matemática.

Aproximándose desde el método NT10, Rios y Soland (2022) estimaron un promedio de respuestas rápidas por estudiante en el dominio de ciencias de PISA 2018 de aproximadamente 6%, tanto para la muestra total como para Uruguay. En Aristas, aplicando un criterio similar, se estima un promedio de 1,4% en lectura y de 3,6% en matemática.

Es interesante destacar que, a diferencia de lo que se observa en otros estudios (Soland, 2018), en Aristas la frecuencia de respuestas rápidas es mayor en matemática que en lectura, lo que podría indicar diferencias en la dificultad percibida o en la estrategia de respuesta de los estudiantes en cada área. Podemos suponer que los menores niveles de esfuerzo observados en la prueba de matemática podrían estar influenciados por el bajo desempeño de los estudiantes, reflejado en los menores niveles de acierto en esta prueba en comparación con la de lectura.

[21] No se mencionan los resultados para lectura dado que la prueba de lectura de PISA 2018 fue adaptativa. En este sentido, el informe de PISA 2018 señala que los ítems fueron asignados a los estudiantes en parte debido a su desempeño anterior en la prueba, lo que impacta en que no se distribuyan uniformemente en su presentación, impidiendo la comparación entre países (OCDE, 2019).

3. **Relación entre el esfuerzo al responder la prueba y las características sociodemográficas y habilidades socioemocionales de los estudiantes**

Se observa que, tal como ha sido confirmado en reiteradas investigaciones en la temática (Avvisati et al., 2024; Rios y Guo, 2020; Soland, 2018; Zamarro, 2021), ser varón incide negativamente en el nivel de esfuerzo, relación que en este estudio se confirmó en ambas áreas.

Asimismo, se observa que en todas las áreas la extraedad incide negativamente en el nivel de esfuerzo, constituyendo el factor con más influencia dentro de las variables sociodemográficas. Vinculado a esto, la bibliografía indica que la adivinación rápida es más alta entre los estudiantes que abandonan sus estudios (Soland y Kuhfeldb, 2019). En efecto, la repetición y el rezago están fuertemente relacionados a la deserción escolar.

Por otro lado, se observa un efecto positivo del nivel socioeconómico del centro en el esfuerzo. Una vez controlada esta característica, el nivel socioeconómico de la familia pierde relevancia al respecto.

La investigación relativa al vínculo entre las habilidades emocionales y la respuesta rápida aún es escasa. Soland y Kuhfeldb (2019) llevaron a cabo un estudio longitudinal con estudiantes de primaria y media en el que concluyeron que la respuesta rápida presenta características de un "estado" (*state-like*) más que de un rasgo duradero en el tiempo (*trait-like*). Como tal, puede ser altamente influenciado por características del ambiente (como el clima de la clase). No obstante, también encontró una asociación entre la respuesta rápida y ciertas habilidades socioemocionales. En particular, la autogestión académica y la autoeficacia fueron dos factores que aparecieron asociados a la respuesta rápida: los estudiantes con mayor presencia de respuestas rápidas informaron una autogestión académica y una autoeficacia mucho menores que las de sus pares sin respuestas rápidas.

En este estudio se exploró la relación del bajo esfuerzo con las habilidades socioemocionales de los estudiantes y se encontró que existe asociación significativa con varias de ellas. La habilidad con asociación más fuerte es la perseverancia académica en matemática, con un coeficiente muy superior a todas las otras habilidades. A continuación, le siguen la autoeficacia académica y la autorregulación metacognitiva, con una relación positiva en ambas áreas, y la motivación intrínseca, significativa solo en lectura. Por otro lado, las habilidades de relacionamiento tienen una relación negativa con el esfuerzo. Por último, las conductas externalizantes muestran una asociación negativa importante, sobre todo en matemática. Al contrario, las conductas internalizantes muestran una pequeña asociación positiva en lectura.

4. **Respecto a las consecuencias del bajo esfuerzo en los resultados de las evaluaciones y posibles abordajes**

Como se ha mencionado, el bajo esfuerzo en la realización de pruebas puede afectar negativamente la confiabilidad y validez de las medidas de habilidad y puede llevar a subestimar el desempeño de algunos estudiantes.

Para abordar esta situación, una posible estrategia es la exclusión de respuestas identificadas con bajo esfuerzo, ya sea a nivel de la respuesta al ítem o del estudiante[22]. Sin embargo, esta decisión conlleva implicaciones metodológicas, ya que reduce el tamaño de la muestra y puede introducir sesgo. Como se ha observado, el bajo esfuerzo no ocurre de manera aleatoria, sino que está asociado a ciertas características de los estudiantes, lo que podría afectar la representatividad de los datos y la interpretación de los resultados.

Otra alternativa para abordar esta cuestión es ajustar las estimaciones de habilidad a posteriori utilizando información sobre el nivel de esfuerzo. Esto puede realizarse mediante ajustes *post hoc* basados en regresiones o incorporando el esfuerzo en la modelización de los puntajes a través de la teoría de respuesta al ítem (Avvisati et al., 2024). Sin embargo, se debe tener en cuenta que este tipo de decisiones afectan los resultados y pueden introducir un nivel indeseado de arbitrariedad en la construcción de las puntuaciones de las pruebas (Avvisati et al., 2024).

Considerando los antecedentes revisados, parecería que contar con cierta cantidad de respuestas realizadas con bajo esfuerzo es una realidad inevitable, incluso más en el marco de las evaluaciones de bajo impacto, sin consecuencias directas para los estudiantes. Como se ha señalado, los porcentajes hallados son similares a los de otras evaluaciones educativas. En este contexto, y más allá de las posibilidades señaladas, la medida más efectiva para abordar el bajo esfuerzo es la prevención, administrando pruebas e ítems que no sean desmotivadores (Avvisati et al., 2024) y en un clima de aplicación adecuado.

En el caso de Aristas, la presencia de un aplicador externo durante la realización de la prueba ha jugado un rol clave en este sentido, ya que, además de garantizar el cumplimiento de los procedimientos estandarizados de aplicación, puede intervenir oportunamente para promover un ambiente adecuado, reforzar instrucciones y fomentar el compromiso de los estudiantes.

Este estudio permitió realizar un primer acercamiento a la detección estadística del bajo esfuerzo en Aristas. Si bien en este trabajo el análisis hizo énfasis en el punto de vista de los alumnos, los indicadores construidos serán utilizados en el futuro para fortalecer los procesos a nivel de ítems, analizando el esfuerzo según dificultad, tipo de ítem, longitud, orden en la prueba, etc., aportando de esta manera evidencia para la compresión de este comportamiento, así como información para mejorar el diseño de las evaluaciones.

En este sentido, es importante seguir monitoreando los tiempos de respuesta en las distintas ediciones de Artistas para analizar la estabilidad del comportamiento y detectar potenciales anomalías que afecten los resultados. También para ampliar el estudio relativo al compromiso, la motivación y el esfuerzo de los estudiantes, incorporando dichas mediciones a los análisis, contribuyendo a una mejor compresión de los logros educativos.

[22] Para calibrar los ítems de las pruebas Aristas se depuran respuestas realizadas en tiempos breves o que resultan inconsistentes.

ANEXO

GRÁFICO A.1
DISTRIBUCIÓN DE LOS UMBRALES POR ÍTEM PARA CADA PRUEBA, SEGÚN MÉTODO

Fuente: elaboración propia a partir de datos de Aristas Media 2022.

VALIDACIÓN DE LOS MÉTODOS UTILIZADOS

CRITERIOS DE VALIDACIÓN

Como criterios de validación de las medidas de esfuerzo obtenidas por los diferentes métodos, se analiza conjuntamente:

1. El porcentaje global de acierto de los estudiantes con bajo esfuerzo, utilizando dos valores del índice RTE para definir el bajo esfuerzo: 1) valores inferiores a 0,9 (estudiantes con más del 10% de los ítems de adivinación rápida) y 2) valores inferiores a 0,7 (más del 30% de los ítems de adivinación rápida)[23]. Se espera que los resultados de estos estudiantes se comporten como respuestas elegidas al azar y que obtengan un porcentaje de acierto inferior al de sus pares que realizaron la prueba con esfuerzo.

2. La variación del porcentaje de acierto según nivel de esfuerzo (punto 1), por nivel socioeconómico. Dado que el nivel socioeconómico es una de las variables mayormente asociadas al desempeño, se espera que los resultados de los estudiantes con mayor esfuerzo se correlacionen con el nivel socioeconómico. En contraste, quienes realizaron la prueba con bajo esfuerzo deberían mostrar porcentajes de acierto similares a los obtenidos por azar, independientemente de su nivel socioeconómico.

3. La correlación de Spearman[24] entre el valor del índice de esfuerzo con respecto al tiempo total dedicado a la prueba y a la habilidad del estudiante. Se espera observar una asociación significativa entre el índice de esfuerzo y estas variables, indicando que los estudiantes que más se esfuerzan dedican más tiempo a la prueba y tienen mayor habilidad estimada.

4. La influencia del esfuerzo en la relación entre variables que teóricamente tienen alta correlación. Se analiza el cambio en la correlación entre la habilidad y el índice socioeconómico del estudiante al depurar la muestra según el índice de esfuerzo. Se espera un aumento en la correlación de dichas variables al eliminar las observaciones de bajo esfuerzo, bajo el supuesto de que los puntajes de habilidad de los casos depurados pueden presentar menor validez.

En la tabla A.1 se puede ver que los porcentajes de acierto de los estudiantes clasificados con bajo esfuerzo a partir del criterio RTE<0,7 se asemejan al nivel de acierto de la elección al azar: alrededor de 0,25[25] y en algunas pruebas levemente superior, cerca de 0,3. Sin embargo, en la clasificación utilizando como punto de corte RTE<0,9 los aciertos superan el nivel del azar. No se observan diferencias sustantivas entre métodos, más que cierta correlación esperable según la exigencia del método, por ejemplo, cuando aumenta el porcentaje del

[23] Estos valores fueron elegidos tomando en cuenta que los estudios en la temática señalan que valores de 0,9 e inferiores indican puntajes sesgados que cuestionan la validez de las estimaciones (Wise, 2015), mientras que 0,7 es un criterio más conservador utilizado para depurar casos (Soland, 2018; Wise y Kong, 2005).
[24] Rios y Guo (2020) usan el coeficiente de correlación de Spearman debido a la no normalidad de la distribución de los tiempos de respuesta.
[25] Los ítems de múltiple opción (CE) considerados en Aristas tienen cuatro opciones de respuesta. Suponiendo que todas tienen la misma probabilidad de respuesta, la probabilidad de acierto si se elige una opción al azar es 25%.

tiempo en el cálculo del umbral normativo (10, 15, 20). También puede observarse que el acierto en los estudiantes clasificados con esfuerzo es similar con independencia del criterio de corte, con un aumento de entre 1 y 2 puntos porcentuales al cambiar el criterio.

TABLA A.1

PORCENTAJE DE RESPUESTAS CORRECTAS EN LOS ESTUDIANTES IDENTIFICADOS CON BAJO ESFUERZO CON CRITERIOS DE CORTE RTE<0,7 Y RTE<0,9, SEGÚN ÁREA Y MÉTODO

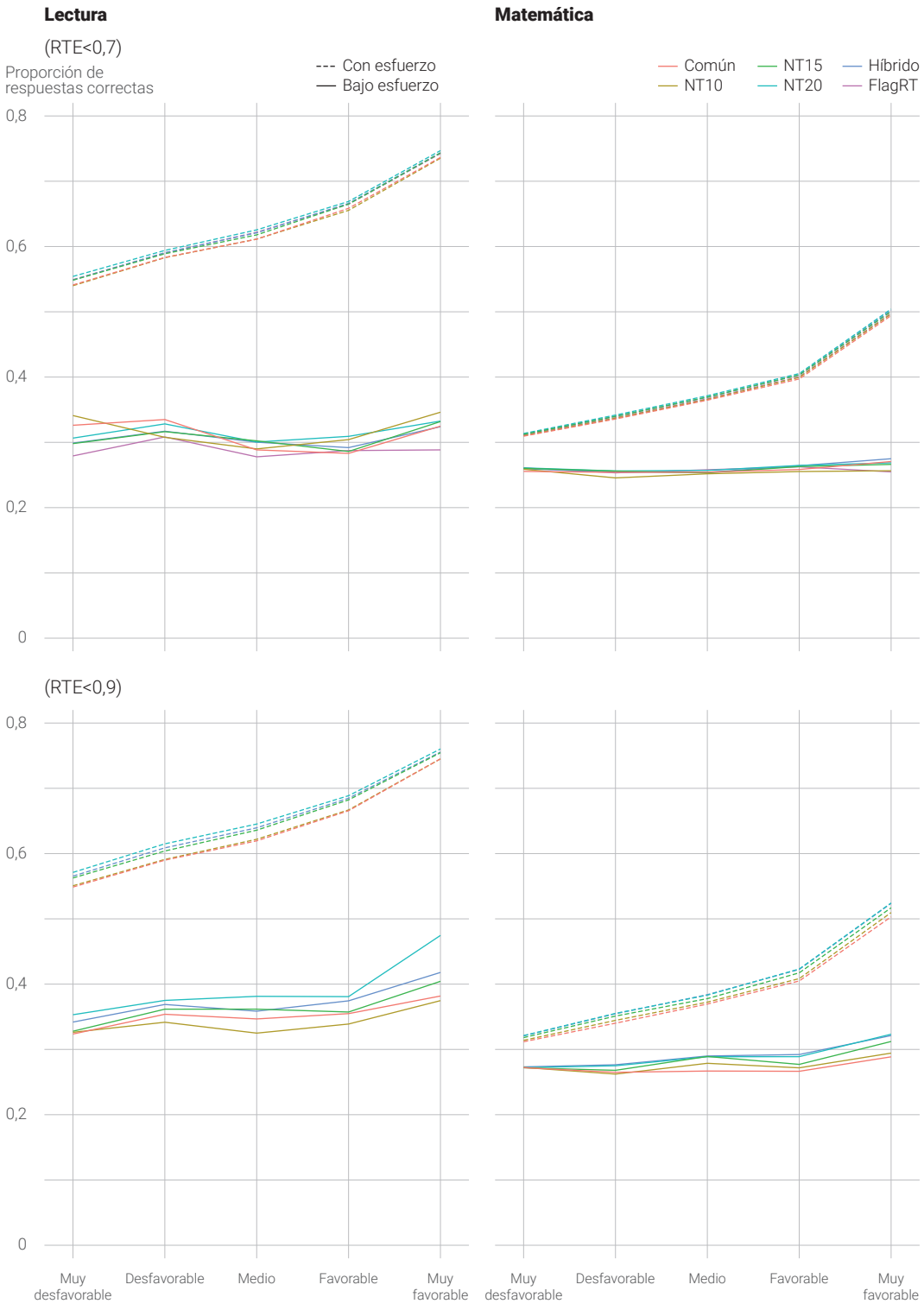
			% de respuestas correctas			
			RTE<0,7	RTE>=0,7	RTE<0,9	RTE>=0,9
9	Lectura	Umbral común	0,31	0,62	0,35	0,63
		NT10	0,32	0,62	0,34	0,63
		NT15	0,31	0,63	0,36	0,64
		NT20	0,31	0,63	0,38	0,65
		Híbrido	0,30	0,63	0,37	0,65
	Matemática	Umbral común	0,25	0,38	0,27	0,38
		NT10	0,25	0,38	0,27	0,39
		NT15	0,26	0,38	0,28	0,4
		NT20	0,26	0,39	0,28	0,4
		Híbrido	0,26	0,38	0,29	0,4

Fuente: elaboración propia a partir de datos de Aristas Media 2022.

Cuando se realiza el análisis por nivel socioeconómico (gráfico A.2), se observa que los resultados de los estudiantes que realizaron la prueba con esfuerzo se asocian positivamente con este: cuanto más favorable el nivel socioeconómico, mayor proporción de aciertos. Al contrario, aquellos estudiantes identificados con bajo esfuerzo tuvieron niveles de acierto cercanos a los de la respuesta al azar, independientemente del contexto. El nivel de acierto de los estudiantes clasificados con bajo esfuerzo se acerca más al azar y es menos dependiente del nivel socioeconómico con criterio RTE<0,7, con respecto al corte RTE<0,9, donde se observa porcentajes de acierto superiores, sobre todo en lectura.

GRÁFICO A.2

PROPORCIÓN DE RESPUESTAS CORRECTAS SEGÚN CATEGORÍA DE ESFUERZO Y NIVEL SOCIOECONÓMICO, SEGÚN MÉTODO DE IDENTIFICACIÓN Y ÁREA



Fuente: elaboración propia a partir de datos de Aristas Media 2022.
Nota: las etiquetas señalan la cantidad de casos clasificados con esfuerzo y bajo esfuerzo en promedio para todos los métodos.

Por otra parte, la correlación del esfuerzo con el tiempo total de respuesta y la habilidad (tabla A.2) es mayor para los índices de esfuerzo que resultan de los métodos NT15, NT20 e híbrido (sobre todo los dos últimos), y más débil en los índices de los métodos más conservadores (umbral común y NT10).

TABLA A.2
CORRELACIÓN DE SPEARMAN ENTRE EL VALOR DEL ÍNDICE DE ESFUERZO CON RESPECTO AL TIEMPO TOTAL DEDICADO A LA PRUEBA Y LA HABILIDAD DEL ESTUDIANTE, SEGÚN MÉTODO DE IDENTIFICACIÓN Y ÁREA

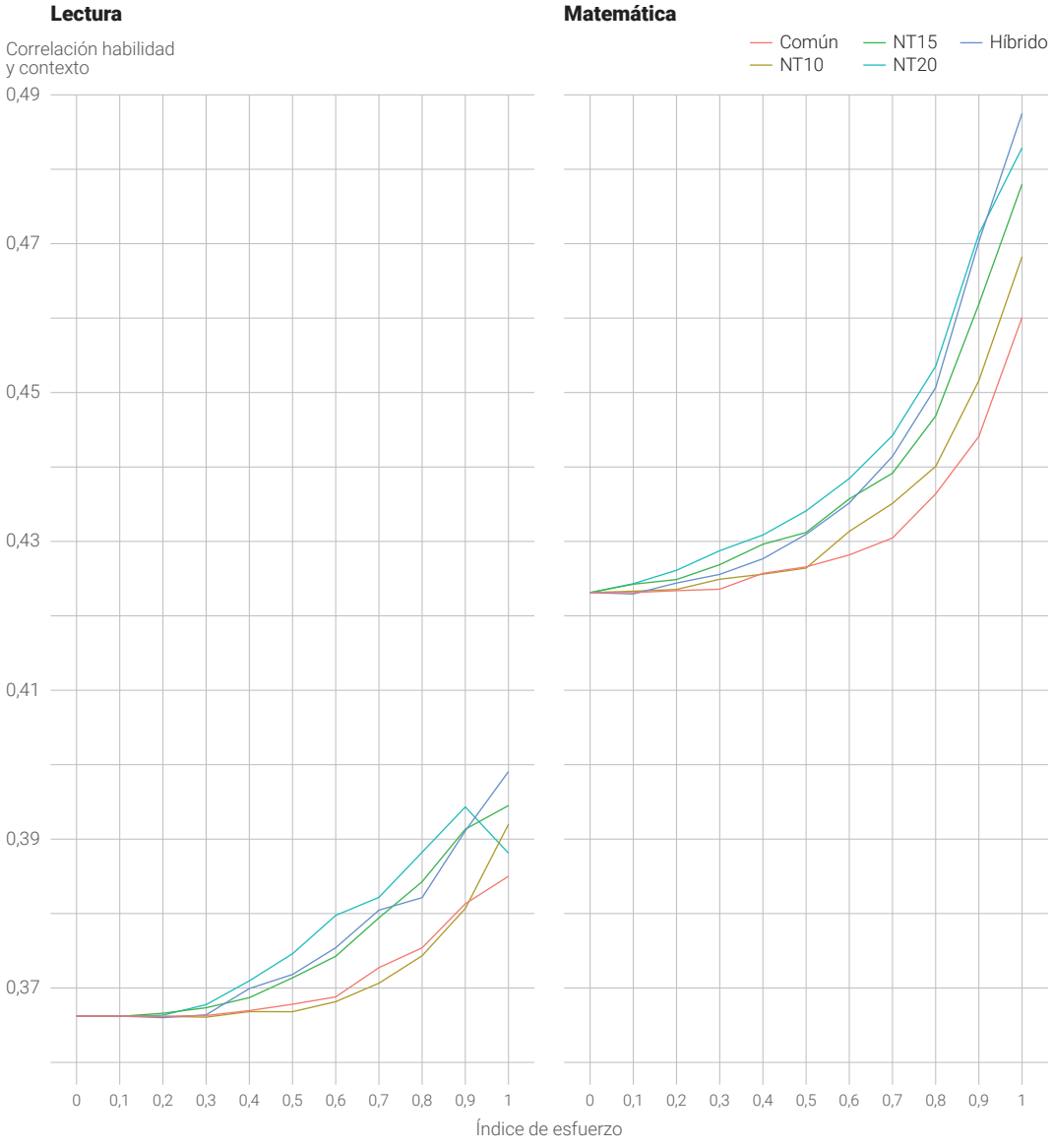
			Correlaciones de Spearman	
			Tiempo total	Habilidad (theta)
9	Lectura	Umbral común	0,26	0,31
		NT10	0,32	0,38
		NT15	0,40	0,42
		NT20	0,45	0,38
		Híbrido	0,41	0,42
	Matemática	Umbral común	0,35	0,26
		NT10	0,41	0,31
		NT15	0,45	0,35
		NT20	0,50	0,37
		Híbrido	0,50	0,35

Fuente: elaboración propia a partir de datos de Aristas Media 2022.

En cuanto al efecto de las respuestas de bajo esfuerzo en la correlación de la habilidad y el contexto, se observa que depurar observaciones con menores valores en los índices de esfuerzo aumenta la correlación entre la habilidad y el índice socioeconómico, consistentemente con la dirección esperada (gráfico A.3). Si se comparan los métodos, el híbrido y NT20 muestran mayor sensibilidad, en tanto la correlación aumenta más pronunciadamente al depurar por dicho criterio, más en matemática. También se observa que los métodos se ordenan según exigencia, siendo menos sensibles los considerados más conservadores (umbral común y NT10).

GRÁFICO A.3

CAMBIO EN LA CORRELACIÓN DE PEARSON ENTRE LA HABILIDAD Y EL ÍNDICE DE NIVEL SOCIOECONÓMICO DEL ESTUDIANTE AL DEPURAR LOS CASOS DE LA MUESTRA CON ÍNDICE DE ESFUERZO MENOR O IGUAL AL VALOR DEL EJE X, SEGÚN MÉTODO DE IDENTIFICACIÓN Y ÁREA



Fuente: elaboración propia a partir de datos de Aristas Media 2022.
Nota: la correlación sin depurar y al depurar por método RT cuantil 50 es de 0,42 a 0,43 en lectura de tercero y de 0,36 a 0,43 en matemática.

TABLA A.3

DESCRIPCIÓN DE LAS HABILIDADES SOCIOEMOCIONALES INCLUIDAS EN EL MODELO MULTINIVEL

Dimensión	Subdimensión	Definición
Motivación y autorregulación del aprendizaje: habilidades con foco en las metas académicas	Autorregulación metacognitiva	Conciencia y control de actividades cognitivas a través de la planificación, el monitoreo y la corrección continua de las actividades cognitivas durante la ejecución de una tarea.
	Motivación intrínseca	Participación en una tarea como un fin en sí mismo, por razones tales como el desafío, la curiosidad o el dominio de una tarea o materia.
	Perseverancia académica	Compromiso con las tareas académicas, foco y persistencia en la persecución de metas académicas, a pesar de obstáculos, dificultades y distracciones.
	Autoeficacia académica	Autovaloración de las habilidades y aptitudes para dominar una tarea académica.
	Valoración de la tarea	Percepciones de los estudiantes sobre la importancia e interés en las tareas.
Habilidades interpersonales: habilidades para la interacción social constructiva	Empatía	Capacidad de entender y compartir el estado emocional de otros, y responder de forma compatible con él a través de la toma de perspectiva, el reconocimiento de emociones y su contexto.
	Habilidades de relacionamiento	Habilidades de comportamiento, socialmente aceptables, que permiten interactuar de forma efectiva con otros.
Habilidades intrapersonales: habilidades para el manejo de las propias emociones y reacciones	Regulación emocional	Estrategias cognitivas para el manejo de la información emocional interna y la regulación de la expresión emocional. Las estrategias remiten a recursos como la capacidad de desviar la atención, de tomar perspectiva o reformular la reacción emocional.
	Autocontrol	Habilidad para controlar reacciones impulsivas frente a situaciones tanto positivas como negativas con el objetivo de cumplir obligaciones y metas a corto plazo. Implica, por lo tanto, la capacidad de cambiar respuestas y evitar conductas indeseables según un contexto determinado.
Conductas de riesgo: remiten a afecciones psicológicas, caracterizadas por problemas emocionales, conductuales y sociales	Conductas internalizantes	Problemas de conducta relacionados a manifestaciones de comportamientos ansiosos, depresivos y problemas somáticos.
	Conductas externalizantes	Problemas de conducta relacionados con hiperactividad, agresividad y conductas oposicionistas.

TABLA A.4

RESULTADOS DEL MODELO MULTINIVEL PARA LAS PRUEBA DE LECTURA Y MATEMÁTICA DE TERCERO

	Lectura			Matemática		
	Valor estimado	Error estándar	Pr(>|t|)	Valor estimado	Error estándar	Pr(>|t|)
(Intercepto)	8,63	0,14	0,000	7,95	0,19	0,000
Contexto socioeconómico y cultural del centro	0,07	0,02	0,000	0,08	0,02	0,000
Estatus socioeconómico y cultural del estudiantes	-0,02	0,01	0,012	-0,01	0,01	0,258
Género masculino	-0,07	0,02	0,000	-0,14	0,02	0,000
Género otro	-0,06	0,05	0,208	-0,12	0,07	0,082
Extraedad 1+	-0,15	0,02	0,000	-0,21	0,03	0,000
Extraedad 2+	-0,15	0,03	0,000	-0,21	0,05	0,000
Región interior	0,08	0,04	0,030	-0,06	0,05	0,213
Autorregulación metacognitiva	0,03	0,01	0,001	0,04	0,01	0,006
Autoeficacia académica	0,04	0,01	0,000	0,10	0,01	0,000
Valoración de la tarea	-0,07	0,08	0,417	0,03	0,11	0,768
Motivación intrínseca	0,02	0,01	0,036	0,01	0,01	0,633
Perseverancia académica	0,04	0,09	0,701	0,61	0,13	0,000
Regulación emocional	0,02	0,01	0,026	0,03	0,01	0,025
Autocontrol	0,01	0,01	0,360	-0,02	0,01	0,191
Habilidades de relacionamiento	-0,02	0,01	0,028	-0,05	0,01	0,001
Empatía	0,01	0,01	0,306	-0,01	0,01	0,547
Conductas externalizantes	-0,10	0,01	0,000	-0,14	0,01	0,000
Conductas internalizantes	0,03	0,01	0,004	0,02	0,01	0,093

Fuente: elaboración propia a partir de datos de Aristas Media 2022
Nota1: en lectura modelo nulo ICC 0,17; R^2 0,17; 9.530 estudiantes, 613 grupos; modelo ICC 0,18; R^2 0,22; 8.875 estudiantes, 612 grupos.
Nota 2: en matemática modelo nulo ICC 0,14; R^2 0,14; 9.358 estudiantes, 611 grupos; modelo ICC 0,13; R^2 0,19; 8.715 estudiantes 611 grupos.

BIBLIOGRAFÍA

AERA, APA y NCME. (2018). *Estándares para pruebas educativas y psicológicas*. Washington D. C.: American Educational Research Association.

Avvisati, F., Buchholz, J., Piacentini, M. y Vargas-Madriz, L. F. (2024). *Item characteristics and test-taker disengagement in PISA* (N.º 312). https://doi.org/10.1787/7abea67b-en

Buchholz, J., Cignetti, M. y Piacentini, M. (2022). *Developing measures of engagement in PISA* (N.º 279). https://doi.org/10.1787/2d9a73ca-en

Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J. y Paek, I. (2016). A New Procedure for Detection of Students' Rapid Guessing Responses Using Response Time. *Applied Measurement in Education*, *29*(3), 173-183. https://doi.org/h10.1080/08957347.2016.1171766

OCDE. (2019). *PISA 2018 Results (Volume I): What Students Know and Can Do*. https://doi.org/10.1787/5f07c754-en

Papanastasiou, E. C. y Michaelides, M. P. (2024). Examining successful and unsuccessful time management through process data: two novel indicators of test-taking behaviors. *Large-scale Assessments in Education*, *12*(3), 1-14. https://doi.org/10.1186/s40536-024-00193-z

Rios, J. A. y Guo, H. (2020). Can Culture Be a Salient Predictor of Test-Taking Engagement? An Analysis of Differential Noneffortful Responding on an International College-Level Assessment of Critical Thinking. *Applied Measurement in Education*, *33*(4), 263-279. https://doi.org/10.1080/08957347.2020.1789141

Rios, J. A. y Soland, J. (2022). An investigation of item, examinee, and country correlates of rapid guessing in PISA. *International Journal of Testing*, *22*(2), 154-184. https://doi.org/10.1080/15305058.2022.2036161

Soland, J. (2018). The Achievement Gap or the Engagement Gap? Investigating the Sensitivity of Gaps Estimates to Test Motivation. *Motivation, Applied Measurement in Education*, *31*(4), 312-323. https://doi.org/10.1080/08957347.2018.1495213

Soland, J. y Kuhfeldb, M. (2019). Do Students Rapidly Guess Repeatedly over Time? A Longitudinal Analysis of Student Test Disengagement, Background, and Attitudes. *Educational Assessment*, *24*(4), 327–342. https://doi.org/10.1080/10627197.2019.1645592

Valdivia, D. S., Rutkowski, L., Rutkowski, D., Canbolat, Y. y Underhill, S. (2023). Test engagement and rapid guessing: Evidence from a large-scale state assessment. *Frontiers in Education*, *8*. https://doi.org/10.3389/feduc.2023.1127644

Wise, S. L. (2015). Effort Analysis: Individual Score Validation of Achievement Test Data. *Applied Measurement in Education*, *28*, 237-252. https://doi.org/10.1080/08957347.2015.1042155

Wise, S. L. y DeMars, C. E. (2006). An Application of Item Response Time: The Effort-Moderated IRT Model. *Journal of Educational Measurement*, *43*(1), 19-38.

Wise, S. L. y Gao, L. (2017). A General Approach to Measuring Test-Taking Effort on Computer-Based Tests. *Applied Measurement in Education*, *30*(4), 343–354. https://doi.org/10.1080/08957347.2017.1353992

Wise, S. L. y Kong, X. (2005). Response Time Effort: A New Measure of Examinee Motivation in Computer-Based Tests. *Applied Measurement in Education*, *18*(2), 163–183.

Wise, S. L. y Ma, L. (2012). Setting Response Time Thresholds for a CAT Item Pool: The Normative Threshold Method. *Meeting of the National Council on Measurement in Education*, 1-24. Vancouver.

Zamarro, G. (2021). *Motivación académica, habilidades no cognitivas y brecha de género en matemáticas y ciencias. El caso de España.* Madrid: Fundación Ramón Areces, Fundación Europea Sociedad y Educación.