

Cómo aumentar la productividad docente con información sobre los estudiantes y las habilidades: Evidencia experimental de Bangladesh*

Sharnic Djaker [†]	Alejandro J. Ganimian [‡]	Noam Angrist [§]	Shwetlena Sabarwal [¶]
Universidad de Oxford	NYU	Universidad de Oxford	Banco Mundial

16 de septiembre de 2026

Resumen

Los docentes pueden no atender los rezagos de aprendizaje porque no saben qué estudiantes y qué habilidades necesitan apoyo. Aunque los exámenes son habituales en todos los sistemas educativos, estas pruebas suelen desaprovecharse. Evaluamos una intervención que brindó a los docentes de matemática de 6.º grado de 312 escuelas de Bangladesh los resultados de dos pruebas breves, informes que identificaban los temas y los estudiantes que necesitaban apoyo, y una clave de respuestas anotada. En la línea de base, los docentes no podían estimar los puntajes de sus estudiantes, su posición en el aula ni su dispersión, y tenían el doble de probabilidades de interactuar con los estudiantes de alto rendimiento que con los de bajo rendimiento. Los informes reorientaron la enseñanza hacia las necesidades que identificaban: los temas y los estudiantes señalados mejoraron más en las evaluaciones de la intervención. En pruebas de cierre independientes, el programa elevó el rendimiento en matemática en 0.23 desvíos estándar (DE) y en lectura en 0.17 DE, con mejoras similares a lo largo de la distribución inicial. También mejoró tres habilidades socioemocionales en 0.10-0.14 DE. Los docentes bajaron sus estimaciones del desempeño de los estudiantes, distribuyeron las interacciones de manera más equitativa y enfatizaron contenidos más fundamentales. A un costo de USD 5.06 por estudiante, estos hallazgos muestran que la información de las evaluaciones que resulta utilizable para la enseñanza y que aprovecha los datos de los exámenes de los sistemas educativos puede ayudar a los docentes a ser sustancialmente más productivos a bajo costo.

*Agradecemos a la Oficina de Asuntos Exteriores, del Commonwealth y de Desarrollo, al Banco Mundial, a la Academia Nacional de Educación, a la Fundación Spencer y a la Fundación Jacobs por financiar este estudio. Agradecemos también a Hunter Gelbach por sus contribuciones en distintas etapas del proyecto. Participantes de seminarios en Chicago, Harvard, NYU y ProLeer comentaron versiones anteriores.

[†]Investigador Científico Senior del What Works Hub for Global Education, Blavatnik School of Government, Universidad de Oxford. Correo electrónico: sharnic.djaker@bsg.ox.ac.uk.

[‡]Profesor Asociado de Psicología Aplicada y Economía, Steinhardt School of Culture, Education, and Human Development, Universidad de Nueva York. Correo electrónico: alejandro.ganimian@nyu.edu.

[§]Director Académico del What Works Hub for Global Education, Blavatnik School of Government, Universidad de Oxford. Correo electrónico: noam.angrist@economics.ox.ac.uk.

[¶]Economista Principal, Educación, Banco Mundial. Correo electrónico: ssabarwal@worldbank.org.

1. Introducción

Un vehículo primordial en el que los gobiernos confían para fomentar el capital humano es la interacción entre estudiantes y docentes en la escuela. Sin embargo, varias fricciones de información pueden limitar la productividad de esas interacciones. Primero, los docentes pueden no saber qué estudiantes necesitan más apoyo; segundo, los docentes también pueden no saber qué habilidades requieren atención. Aunque a menudo existe información sustancial sobre los estudiantes y las habilidades a través de los exámenes regulares en casi todos los sistemas educativos del mundo, estas pruebas suelen desaprovecharse. Esta brecha es relevante: introduce grandes fricciones de información que atenúan la productividad de las interacciones entre docentes y estudiantes, pero que podrían atenderse a bajo costo y a escala aprovechando mejor los exámenes que ya existen en el sistema. Cerrar este tipo de fricciones de información ha resultado importante en el contexto de las interacciones entre padres e hijos (Bergman, 2020; Dizon-Ross, 2019) y podría ser aún más esencial en el contexto de las interacciones entre docentes y estudiantes, lo que ayudaría a explicar por qué, a pesar de las expansiones históricas de la escolarización en todo el mundo, los niveles de aprendizaje de los estudiantes siguen siendo bajos y estancados (Angrist et al., 2021; World Bank, 2018a).

En este trabajo, evaluamos experimentalmente una intervención diseñada para cerrar las fricciones de información en dos márgenes, identificando *estudiantes* y *habilidades* que necesitan apoyo, brindando a los docentes de matemática de 6.º grado de 312 escuelas de Bangladesh resultados accionables de las pruebas. Tomamos una muestra representativa de 312 escuelas secundarias de asociación público-privada (APP)—que matriculan a 8 de cada 10 estudiantes—y las asignamos aleatoriamente a un grupo de tratamiento o de control. En el primero, evaluamos a los estudiantes de 6.º grado al inicio y a mitad del año y, poco después, ofrecimos a los docentes informes que identificaban los temas y los estudiantes que necesitaban apoyo. Los docentes también recibieron la prueba y una explicación de por qué cada respuesta era correcta o incorrecta, para reforzar conceptos clave e identificar las concepciones erróneas subyacentes a los errores frecuentes de los estudiantes. Las escuelas de control no participaron en las evaluaciones ni recibieron los informes ni las guías de errores, de modo que pudimos evaluar su impacto en relación con la enseñanza y el aprendizaje habituales.

Reportamos los resultados en tres segmentos principales. Primero, cuantificamos las fricciones de información entre los docentes y sus consecuencias para la enseñanza. Los docentes no podían estimar el desempeño de sus estudiantes en las pruebas. Evaluamos a los estudiantes en matemática y pedimos a los docentes que estimaran sus puntajes. La correlación entre los puntajes reales y los estimados fue de 0.23, de modo que las estimaciones explicaron solo el 5 % de la variación en los puntajes. Los docentes no parecían saber qué estudiantes necesitaban más apoyo: la correlación entre las posiciones reales de los estudiantes dentro del aula y las implícitas en las estimaciones fue de 0.16, lo que significa que los ordenamientos de

los docentes predijeron menos del 3 % de la variación en los reales. Sus errores fueron grandes en relación con la distribución del aula: sobreestimaron los puntajes de sus estudiantes de menor rendimiento—quienes más necesitan su apoyo—en 6.9 puntos porcentuales (p.p.), lo que equivale al 50 % del desvío estándar (DE) dentro del aula en los puntajes. Los docentes tampoco sabían cuánto variaban sus estudiantes en su rendimiento: la correlación entre los DE dentro del aula de los puntajes reales y las estimaciones de los docentes fue de -0.01. Esto sugiere que los docentes pueden atender insuficientemente las necesidades de sus estudiantes en parte porque desconocen cuántos están en dificultades o en qué medida lo están.

Los docentes tenían pocas oportunidades de actualizar sus creencias previas incorrectas sobre las habilidades de sus estudiantes. Observamos clases y recolectamos datos detallados sobre cada intercambio verbal que un docente tuvo con un estudiante individual o con un grupo pequeño. En la clase típica de matemática de 40 minutos, hubo solo 8.5 de esos intercambios; la mayor parte de la clase se dedicó a la enseñanza al grupo completo. El estudiante promedio interactuó solo 0.24 veces con su docente, pero incluso esos escasos intercambios se distribuyeron de manera desigual: los estudiantes de menor rendimiento interactuaron la mitad de las veces que sus pares de mayor rendimiento; de hecho, 9 de cada 10 estudiantes de bajo rendimiento no interactuaron con sus docentes en absoluto. Estas diferencias las generaron los docentes (no los estudiantes), que iniciaron 7 de cada 10 intercambios.

Segundo, reportamos los efectos de la intervención, que fue diseñada para cerrar estas brechas mediante información tanto sobre los estudiantes como sobre las habilidades. Los docentes parecieron responder a los informes. En las escuelas de tratamiento, los estudiantes mejoraron 14.5 p.p. más en los temas (por ejemplo, medición y geometría) señalados por bajo desempeño que en los temas no señalados. También mejoraron más en los subtemas señalados (por ejemplo, aplicar propiedades geométricas para hallar un ángulo faltante) en 19.4 p.p. que en los subtemas no señalados. Incluso el color con el que mostramos el desempeño por ítem pareció marcar una diferencia: los estudiantes mejoraron 34.7 p.p. más en los ítems mostrados en rojo (bajo desempeño) o amarillo (medio) que en los ítems mostrados en verde (alto). Los estudiantes mencionados por nombre en los informes también mejoraron 9.9 p.p. más, y la magnitud de sus mejoras aumentó con la cantidad de temas en los que fueron señalados por bajo desempeño. Estos estudiantes incluso mejoraron 32.9 p.p. más en los temas por los que fueron señalados. Estos patrones implican que los docentes reorientaron la enseñanza para atender los desafíos específicos destacados por los informes.

La intervención elevó el rendimiento más allá de las pruebas utilizadas para producir los informes. Comparamos a los estudiantes de tratamiento y de control en pruebas de cierre independientes y encontramos que la intervención aumentó el rendimiento en matemática en 0.23 DE, con impacto en todos los temas y habilidades evaluados. Las mejoras fueron amplias, con impactos de magnitudes similares a lo largo de la distribución inicial. Los estudiantes

nombrados en los informes superaron a sus pares de control que habrían sido nombrados, pero también lo hicieron los estudiantes no nombrados en márgenes similares. La intervención también aumentó los puntajes de las pruebas de lectura en 0.17 DE, aun cuando se centró en matemática, y no podemos descartar que los efectos hayan sido iguales a lo largo de la distribución de rendimiento. Las mejoras se concentraron en los ítems de vocabulario y comprensión—dos habilidades que se ejercitan en los problemas con enunciado y que son aplicables a todas las materias. En conjunto, estos impactos muestran que los cambios en la enseñanza que los docentes adoptaron en respuesta a los informes beneficiaron a todos los estudiantes de sus aulas, no solo a los estudiantes de bajo rendimiento.

La intervención también impactó algunas medidas del esfuerzo y la motivación de los estudiantes. No tuvo efectos sobre la asistencia, medida tanto por los registros escolares como por la proporción de estudiantes matriculados presentes en los días de recolección de datos. Sin embargo, el estudiante de tratamiento promedio superó a su contraparte de control en autoeficacia (0.14 DE), mentalidad de crecimiento (0.11 DE) y orientación a metas de tarea (0.10 DE). Los estudiantes de bajo rendimiento registraron mejoras estadísticamente significativas en la asistencia basada en registros y en las tres medidas socioemocionales. Sin embargo, con dos excepciones, no podemos descartar que los efectos hayan sido iguales a los de sus pares de rendimiento medio y alto. Consideramos estos efectos como evidencia adicional de que la intervención produjo mejoras generalizadas entre los estudiantes.

Tercero, en cuanto a los mecanismos, encontramos cambios mayores en las habilidades que los docentes enseñan (al grupo completo) que en la focalización en estudiantes específicos. Los docentes ajustaron a la baja sus estimaciones del puntaje del estudiante promedio en 0.47 DE, y no podemos rechazar que hayan ajustado en una magnitud similar para los estudiantes a lo largo de la distribución de rendimiento.¹ De manera consistente con este cambio uniforme en sus estimaciones, los docentes comenzaron a distribuir las interacciones de manera más equitativa: la proporción de las interacciones que recibieron los estudiantes de bajo rendimiento aumentó 5 p.p. (del 22 % al 27 %), reflejando la caída de 5 p.p. en la de los estudiantes de alto rendimiento (del 45 % al 40 %). En consecuencia, también reorientaron la enseñanza de temas más avanzados a temas más fundamentales: durante las observaciones, los docentes de tratamiento tuvieron 15.3 p.p. menos de probabilidad que los docentes de control de cubrir datos y 13.6 p.p. más de probabilidad de cubrir número.² En conjunto, estos cambios ayudan a explicar los impactos generalizados sobre las habilidades académicas y socioemocionales.

Nuestra primera contribución es a la literatura sobre información. Trabajos previos muestran que los padres a menudo tienen creencias inexactas sobre las habilidades de sus hijos y que corregirlas puede cambiar sus inversiones y mejorar los resultados de aprendizaje

¹Los informes no hicieron que los docentes fueran más precisos sobre los puntajes individuales, lo que tiene sentido porque identificaban estudiantes y temas que necesitaban apoyo en lugar de dar puntajes.

²Las estimaciones sobre estas medidas son menos precisas porque se basan en una submuestra de aulas.

(Neilson et al., 2015; Bergman, 2020; Dizon-Ross, 2019; Evans and Acosta, 2026). Encontramos que los docentes también tienen creencias inexactas sobre el rendimiento de sus estudiantes y, como los padres, subutilizan la información ya disponible. Estudios previos también advierten que la información por sí sola puede no bastar para producir cambios duraderos en el comportamiento de los docentes (Muralidharan and Sundararaman, 2010). Nuestro estudio—leído junto con otros (Andrabi et al., 2017; Camargo et al., 2018; de Hoyos et al., 2021; de Hoyos et al., 2023; Cilliers and Habyarimana, 2026)—sugiere que cuando la información se usa para cambiar lo que el sistema escolar espera de los docentes, es más probable que eleve su esfuerzo. También documentamos *cómo* ocurre esta respuesta en el aula: los docentes reorientaron la enseñanza hacia los temas identificados como necesitados de apoyo, mejorando a bajo costo el rendimiento estudiantil. Los efectos que encontramos al compartir resultados de evaluaciones sin prescribir un enfoque pedagógico sugieren que la información puede hacer más productiva la autonomía que los docentes ya tienen (Aghion and Tirole, 1997), lo que concuerda con la creciente evidencia (Beg et al., 2022, 2024; Piza et al., 2024; Andrabi et al., 2025; Nourani et al., 2025) de la educación en países de ingresos bajos y medios (PIBM).

Nuestra segunda contribución es a una pregunta clave en la intersección del desarrollo y la educación: si los docentes de los PIBM “enseñan a la cima” de la distribución de rendimiento. La preocupación de que los docentes puedan atender a sus mejores estudiantes a expensas de sus pares (Banerjee and Duflo, 2011) motivó algunas de las evaluaciones más influyentes en educación, incluidas la enseñanza remedial (Banerjee et al., 2007), el agrupamiento por habilidad (Duflo et al., 2011), los campamentos de aprendizaje (Banerjee et al., 2010) y la enseñanza diferenciada (Banerjee et al., 2017). Sin embargo, se ha inferido sobre todo de los incentivos del sistema—como currículos excesivamente ambiciosos (Pritchett and Beatty, 2015) y exámenes de altas consecuencias (Glewwe et al., 2009)—y de la amplia brecha entre los estándares curriculares y la preparación de los estudiantes para la escuela (Muralidharan et al., 2019). Somos de los primeros en aportar evidencia *directa* de que la enseñanza habitual sí se centra en temas avanzados y en estudiantes de alto rendimiento.³ Nuestros resultados sugieren que la enseñanza puede estar desalineada con las necesidades de los estudiantes también porque los docentes desconocen su rendimiento. La información de bajo costo basada en evaluaciones de estudiantes puede reorientar la enseñanza y mejorar la productividad de los docentes sin cambios mayores en las estructuras o los incentivos del sistema educativo.⁴

³Nuestras medidas basadas en observación se apoyan en instrumentos previos (Abadzi, 2009; Bhattacharjea et al., 2011; Bruns and Luque, 2014; Sankar and Linden, 2014; Stallings et al., 2014; World Bank, 2017) que muestran que los docentes dedican la mayor parte de la clase a la enseñanza al grupo completo y usan los mismos materiales con todos los estudiantes.

⁴La correlación entre los puntajes estimados y los reales es consistente con nuestro trabajo previo en India y Bangladesh (Djaker et al., 2024) y con replicaciones en Benín, Nigeria y Gambia (Rodríguez-Segura, 2024), y es un octavo de la magnitud observada en países de ingreso alto (Hoge and Coladarci, 1989; Südkamp et al., 2012; Kaufmann, 2024), lo que sugiere que esta posibilidad merece mayor consideración.

En relación con esto, nuestra tercera contribución es a la investigación sobre la enseñanza diferenciada en la escuela—una de varias intervenciones que se agrupan bajo el término “Enseñar al Nivel Adecuado” (TaRL, por su sigla en inglés). Evaluaciones aleatorizadas en India, Kenia, Ghana, Costa de Marfil y Zambia han encontrado que, cuando se implementa con fidelidad, este enfoque—que combina evaluar a los estudiantes, agruparlos por desempeño y asignar distintas actividades a cada grupo—mejora las habilidades fundamentales (Dufflo et al., 2011; Banerjee et al., 2017; Dufflo et al., 2024; de Barros and Lubozha, 2026; Wolf et al., 2026). Nuestra intervención arroja luz sobre uno de sus mecanismos. Brindar a los docentes resultados de evaluaciones—sin agrupar a los estudiantes ni asignar actividades separadas—los llevó a ajustar a la baja sus creencias sobre el rendimiento de los estudiantes, a distribuir las interacciones de manera más equitativa y a enseñar temas más fundamentales, al tiempo que elevó el aprendizaje a lo largo de la distribución inicial. Estos hallazgos sugieren que actualizar las creencias de los docentes puede dar cuenta de una proporción no trivial de los efectos de la enseñanza diferenciada, al cambiar cómo los docentes asignan su atención y cómo calibran la enseñanza en el aula para grupos académicamente diversos de estudiantes.

Finalmente, contribuimos a la investigación sobre cómo mejorar el capital humano de manera costo-efectiva en los PIBM—una prioridad central dados los niveles de aprendizaje persistentemente bajos en todo el mundo (Kremer et al., 2013; World Bank, 2018b; Angrist et al., 2021). Aunque los exámenes escolares son habituales en casi todos los sistemas educativos, sus resultados suelen desaprovecharse. Nuestros hallazgos sugieren que convertir los resultados de las evaluaciones en información accionable para los docentes podría mejorar el aprendizaje de manera más costo-efectiva que las alternativas principales. Nuestra intervención generó 5.68 Años de Escolaridad Ajustados por Aprendizaje (LAYS, por su sigla en inglés) en matemática y 4.20 LAYS en lectura por cada USD 100, superando las estimaciones de dos enfoques muy extendidos en la educación global: la pedagogía estructurada y TaRL (GEEAP, 2020; Angrist et al., 2025). Esto sugiere que los sistemas educativos podrían aprovechar mejor las evaluaciones existentes para mejorar el aprendizaje costo-efectivamente y a escala.

2. Experimento

2.1. Contexto

Realizamos este estudio en Bangladesh, un país de ingreso mediano bajo del sur de Asia cuya población es comparable en tamaño a las de Brasil, Rusia y Etiopía (World Bank, 2026). En 2021, su ministerio de educación solicitó apoyo al Banco Mundial para desarrollar una intervención de evaluación formativa que pudiera implementarse a escala nacional. Insertamos un piloto de esa intervención dentro de una evaluación aleatorizada de gran escala para estimar su impacto e identificar los ajustes que pudieran ser necesarios. Dado que Bangladesh tiene

el tercer sistema de educación secundaria más grande del sur de Asia—con 15.1 millones de estudiantes matriculados (UIS, 2026)—esperamos que las lecciones de nuestro estudio resulten útiles tanto para este país como para sus vecinos de la región.

Como en muchos PIBM, Bangladesh ha expandido rápidamente el acceso a la escolarización en las últimas décadas. La educación primaria abarca de 1.º a 5.º grado y es obligatoria; la educación secundaria abarca de 6.º a 12.º grado y no lo es (NCTB, 2013). La matrícula neta en el primer ciclo de secundaria se ha más que triplicado desde 1990, hasta alcanzar el 72 % (BANBEIS, 2023). La matrícula ha superado la capacidad del país para construir escuelas y contratar docentes, de modo que el ritmo acelerado de expansión ha dado lugar a aulas más numerosas y académicamente más diversas. La cantidad de estudiantes de secundaria por docente pasó de 27 en 1990 a 35 en 2018 (UIS, 2026). Y 8 de cada 10 estudiantes de secundaria son estudiantes de primera generación—es decir, tienen padres que no han completado este nivel (Barro and Lee, 2026), y están por lo tanto menos preparados para ayudar a sus hijos a transitarlo. Las escuelas carecen de los recursos necesarios para apoyar a estos estudiantes: el gasto por estudiante en la educación secundaria es de USD 371, muy por debajo de los promedios del sur de Asia (USD 915) y de los PIBM (USD 924) (UNESCO, 2023).

A la luz de este contexto, quizás no sorprenda que los resultados de aprendizaje sigan siendo bajos y desiguales. En 2014, solo el 40 % de los estudiantes de 5.º grado podía resolver problemas de resta de dos dígitos con acarreo y el 31 % podía resolver problemas de división de tres dígitos por un dígito (BRAC, 2015). Además, solo el 34 % de los estudiantes cuyas madres no tenían escolaridad obtuvo un promedio de calificaciones de 4 o más, en comparación con el 80 % de aquellos cuyas madres tenían 10 o más años de escolaridad.

Varias características del sistema bangladesí pueden dar a los docentes razones para centrarse en los estudiantes de mayor rendimiento. Primero, dado que la secundaria no es obligatoria, los docentes pueden anticipar que los estudiantes de menor rendimiento tienen más probabilidades de abandonar e invertir menos en ellos (Banerjee and Duflo, 2011; Duflo et al., 2011). Segundo, hasta hace poco, el currículo se parecía al de otros PIBM en su énfasis en la memorización de hechos (NCTB, 2021), lo que potencialmente dejaba poco tiempo a los docentes para verificar si los estudiantes comprendían el material (Glewwe et al., 2009; Banerjee and Duflo, 2011; Muralidharan and Zieleniak, 2014; Pritchett and Beatty, 2015; Muralidharan et al., 2019; Rodríguez-Segura and Mbiti, 2022). Tercero, los estudiantes de 6.º grado rinden la Evaluación Nacional de Estudiantes de Secundaria (NASS, por su sigla en inglés)—una evaluación de matemática, bangla e inglés (DSHE, 2019)—lo que puede llevar a sus docentes a invertir más esfuerzo (Glewwe et al., 2010; Duflo et al., 2011; Gilligan et al., 2022) en los estudiantes con mayores probabilidades de aprobar los exámenes.

2.2. Muestreo

Muestreamos 312 escuelas secundarias de asociación público-privada (APP) en tres divisiones de Bangladesh: Chattogram, Mymensingh y Rajshahi.⁵ Nos centramos en las escuelas APP porque matriculan a 8 de cada 10 estudiantes de secundaria (DSHE, 2019). Tomamos nuestra muestra de la siguiente manera. Primero, muestreamos aleatoriamente tres de las ocho divisiones del país, estratificando nuestra selección por terciles de los puntajes promedio de la NASS de las divisiones (es decir, una división por tercil). En el paso siguiente, seleccionamos aleatoriamente cuatro distritos (de un promedio de ocho) por división, estratificando por la mediana de los puntajes promedio de la NASS de los distritos (es decir, dos distritos por encima y dos por debajo de la mediana). Luego, excluimos las escuelas no APP y las madrasas (que en conjunto representan el 19% de la matrícula secundaria) y las escuelas con secciones de 6.º grado muy pequeñas (por debajo del percentil 5) o muy grandes (por encima del percentil 75). Por último, entre las escuelas restantes, muestreamos 26 por distrito (es decir, 312 en total) y, dentro de las escuelas muestreadas, seleccionamos aleatoriamente una sección de 6.º grado por escuela y hasta 20 estudiantes por sección, para un total de 5,935 estudiantes. Estos estudiantes constituyen nuestra muestra de evaluación de línea de base, a la que buscamos seguir a lo largo de todas las rondas de recolección de datos. Los estudiantes de las escuelas de tratamiento también rindieron las evaluaciones de la intervención y por lo tanto podían aparecer por nombre en los informes para los docentes.

Las escuelas de la muestra tuvieron un desempeño en términos generales similar al de las escuelas fuera de la muestra en los puntajes de matemática y bangla de la NASS, tanto a nivel nacional como dentro de las divisiones muestreadas (Tabla A.1 en el Apéndice A). Esto sugiere que las lecciones que extraemos de este experimento probablemente se extiendan no solo a los aproximadamente 3.9 millones de estudiantes de secundaria matriculados en las divisiones muestreadas, sino también a los aproximadamente 9.4 millones matriculados en todo el país.

2.3. Aleatorización

Asignamos aleatoriamente las 312 escuelas a la intervención de evaluación formativa descrita abajo o a un grupo de control con la práctica habitual, estratificando por distrito.⁶

La asignación aleatoria creó grupos de estudiantes comparables en la mayoría de las características. El estudiante de control promedio tenía 12 años; el 61 % eran mujeres, el 94 % hablaba bangla en el hogar, y sus puntajes promedio de matemática, lectura e inteligencia fluida fueron estandarizados a cero. Los estudiantes de tratamiento tuvieron una probabilidad

⁵La Figura A.1 del apéndice muestra la distribución geográfica de las escuelas muestreadas.

⁶Este estudio se realizó como parte de una asociación institucional entre el Banco Mundial y el Gobierno de Bangladesh, en el marco de la cual también estudiamos una intervención en la que además se pidió a los docentes que evaluaran a los estudiantes. Su resultado se reporta en un trabajo complementario (Djaker, 2023).

levemente mayor de ser mujeres, menor de tener una madre que completó la secundaria, mayores activos del hogar y menor rendimiento en matemática (Tabla A.2). Sospechamos que la diferencia en matemática refleja el orden diferencial en que se administraron las pruebas en la línea de base entre los grupos, más que diferencias subyacentes no observables.⁷ Sin embargo, en tanto estas diferencias iniciales reflejen diferencias reales en el rendimiento previo, sugieren que los estudiantes de tratamiento comenzaron algo por detrás de los de control; en todo caso, esto haría conservadoras las estimaciones no ajustadas del impacto del programa.

La asignación aleatoria también creó grupos de docentes similares. El docente de control promedio tenía 42 años y 14 años de experiencia docente; el 12 % eran mujeres; los docentes de tratamiento tuvieron menor probabilidad de haber recibido formación inicial (Tabla A.3). Dado que realizamos 14 pruebas de balance de docentes, una diferencia estadísticamente significativa es aproximadamente lo que esperaríamos por azar. Una prueba conjunta de todas las covariables docentes listadas no rechaza la igualdad entre los grupos.

2.4. Atrición

La mayoría (77 %) de los estudiantes que participaron en la ronda de línea de base de recolección de datos también participaron en la de cierre. Las tasas de atrición fueron casi idénticas entre los grupos experimentales, y no encontramos evidencia de atrición diferencial respecto de las características de los estudiantes en la línea de base (Tabla A.4).

2.5. Intervención

La intervención constó de tres componentes: dos pruebas breves de matemática (es decir, de 10 ítems de opción múltiple) a lo largo de ocho meses; dos informes breves (es decir, de cinco páginas) basados en esos resultados, entregados a los docentes poco después de cada ronda de evaluación; y una guía que identificaba la respuesta correcta de cada ítem, explicaba cómo llegar a esa respuesta y describía las concepciones erróneas que podrían explicar por qué los estudiantes eligieron respuestas incorrectas. La Figura A.3 muestra el cronograma de sus actividades (panel inferior).⁸ El Apéndice B detalla cada uno de sus componentes.

Hay varios aspectos del diseño de la intervención que buscaron hacerla útil para los docentes. Primero, mantuvimos las pruebas breves para que fueran fáciles y de bajo costo de administrar y para que los docentes pudieran familiarizarse con todos los ítems de la prueba y con el proceso deseado para responderlos. También esperábamos que esto facilitara escalar la intervención, si resultaba exitosa. Esta extensión breve exigió que eligiéramos los temas y las habilidades a evaluar de manera deliberada, por lo que basamos nuestra asignación

⁷Solo los estudiantes de tratamiento rindieron las pruebas de la intervención, previas a las de evaluación.

⁸Debido a los cierres de escuelas provocados por el COVID-19, el año escolar comenzó el 22 de febrero en lugar del 21 de enero, como estaba previsto originalmente.

de ítems en el marco de evaluación de una prueba internacional de matemática de 4.º grado, suponiendo que muchos estudiantes de 6.º grado probablemente estarían desempeñándose por debajo del nivel de su grado.⁹ Segundo, administramos dos rondas de pruebas, con la segunda ronda meses antes del fin del año, para darles un incentivo para mejorar de la primera a la segunda ronda y para demostrarles que los estudiantes de bajo rendimiento efectivamente eran capaces de mejorar.¹⁰ Tercero, los informes identificaban los temas y los estudiantes específicos (por nombre) que necesitaban apoyo, para enfocar la atención de los docentes en ellos.¹¹ En investigaciones previas en India y Bangladesh, habíamos encontrado que los docentes subestimaban la proporción de sus estudiantes que estaba por detrás de las expectativas curriculares y la medida en que lo estaba (Djaker et al., 2024). Queríamos entender si hacer más visibles a los estudiantes de bajo rendimiento haría que los docentes estuvieran más dispuestos a ayudarlos.¹² Cuarto, los resultados se presentaron en formato narrativo para que fueran más fáciles de comprender.¹³ Por último, la guía no solo identificaba las respuestas correctas, sino que también explicaba cómo llegar a ellas y qué podrían estar malinterpretando los estudiantes al seleccionar cada opción incorrecta (diseñada deliberadamente como distractor), para que los resultados de la prueba resultaran informativos para la enseñanza.¹⁴

Estas características se corresponden con una teoría del cambio simple pero coherente: las pruebas buscaban enfocar la atención en las habilidades fundamentales; los informes pretendían impulsar a los docentes a actualizar sus creencias previas sobre los temas y los estudiantes con mayor necesidad de apoyo; la guía debía traducir esa toma de conciencia en decisiones de enseñanza; y el tiempo entre rondas dio a los docentes tiempo y un incentivo para mejorar el desempeño de sus estudiantes de rendimiento más bajo.

⁹Este supuesto se basó tanto en estudios previos en PIBM (Muralidharan et al., 2019; Büchel et al., 2020) como en los resultados de la NASS (véase la sección 2.1). Luego verificamos que los temas, habilidades e ítems evaluados se alinearan con el currículo bangladés y pedimos comentarios a docentes de Teach for Bangladesh.

¹⁰Estos objetivos parecían importantes porque, en una encuesta reciente de docentes en 11 PIBM, muchos reportaron que no creían poder ayudar a los estudiantes si sus padres no se involucraban en su educación, y que no consideraban su responsabilidad enseñar material de grados inferiores (Sabarwal et al., 2022).

¹¹Describimos en detalle cómo se señalaron los temas y los estudiantes en la sección 5.2.1.

¹²Este enfoque difiere de otras intervenciones de información en educación, que comparan el desempeño de las escuelas con el de otras de la zona (Muralidharan and Sundararaman, 2010; Andrabi et al., 2017). El enfoque en Bangladesh se inspiró en evidencia alentadora de Argentina (de Hoyos et al., 2021; de Hoyos et al., 2023), que mostró resultados alentadores de informes que combinaban esa comparación con la información sobre los temas que necesitaban refuerzo.

¹³En un proyecto anterior, habíamos notado que los docentes tenían dificultades para interpretar los gráficos y respondían de manera más positiva a informes presentados en forma narrativa (Banerjee et al., 2018).

¹⁴Trabajos previos (Tatto et al., 2012; Bruns and Luque, 2014; Bold et al., 2017; Ganimian et al., 2026) muestran que los docentes de los PIBM tienen dificultades con habilidades fundamentales que deben enseñar.

3. Datos

Realizamos tres rondas de recolección de datos durante el año escolar 2022: una línea de base previa a la intervención (que incluyó encuestas a estudiantes y docentes, evaluaciones y datos de asistencia), una línea intermedia a mitad del año (que incluyó visitas escolares sin aviso previo para medir la asistencia de los estudiantes) y una línea de cierre al final del año (que incluyó encuestas ampliadas a estudiantes y docentes, evaluaciones de estudiantes, observaciones de aula y datos administrativos sobre la asistencia de los estudiantes). La Figura A.3 muestra el cronograma de las actividades de investigación (panel superior).

3.1. Asistencia de estudiantes y docentes

Medimos la asistencia de estudiantes y docentes utilizando los registros escolares. En la línea intermedia, los encuestadores transcribieron el total de días laborables y de días presentes de enero a julio de cada estudiante muestreado y del docente de matemática seleccionado. En la línea de cierre, recolectaron las mismas medidas para el período de agosto a octubre. Calculamos la asistencia como la cantidad total de días presentes dividida por el total de días laborables de los meses con registros válidos. Estos datos cubren 5,941 estudiantes y 305 docentes en la línea intermedia y 7,039 estudiantes y 311 docentes en la línea de cierre.

3.2. Características y habilidades socioemocionales de los estudiantes

Encuestamos a los estudiantes en la línea de base y en la de cierre. En la línea de base, recolectamos variables demográficas y socioeconómicas que utilizamos para describir la muestra, evaluar el balance entre los grupos experimentales y probar efectos heterogéneos. En la línea de base y en la de cierre, también medimos sus habilidades socioemocionales para ajustar por diferencias preexistentes, probar si los impactos sobre el rendimiento variaban según esas habilidades y estimar impactos sobre las habilidades mismas.

3.3. Características y creencias de los docentes

También encuestamos a los docentes en la línea de base y en la de cierre. En la línea de base, recolectamos variables demográficas, educativas y profesionales para describir la muestra de docentes, evaluar el balance y probar efectos heterogéneos según sus antecedentes. En la línea de base y en la de cierre, también medimos sus creencias y prácticas para ajustar por diferencias preexistentes y probar heterogeneidad, y estimar impactos sobre estos resultados.

3.4. Rendimiento de los estudiantes

Evaluamos a los estudiantes en la línea de base y en la de cierre. En la línea de base, medimos el rendimiento en matemática y lectura y la inteligencia fluida para describir la muestra, evaluar el balance, ajustar por diferencias preexistentes y probar efectos heterogéneos según las habilidades previas. En la línea de cierre, evaluamos matemática para estimar impactos y lectura para probar efectos de derrame. Las pruebas de matemática asignaron los ítems entre temas (número, medición y geometría, y datos) y habilidades (conocer, aplicar y razonar) siguiendo el mismo marco de evaluación que las evaluaciones de la intervención (véase la sección B.1).¹⁵ Aprovechamos los ítems comunes entre la línea de base y la de cierre para mapear ambas rondas de evaluaciones en una escala común mediante un modelo de teoría de respuesta al ítem (TRI) logístico de dos parámetros. Estandarizamos todos los puntajes escalados por TRI con respecto a la distribución del grupo de control.

3.5. Estimaciones de los docentes sobre el rendimiento de sus estudiantes

En un estudio previo, habíamos indagado las estimaciones de los docentes sobre el desempeño de sus estudiantes en pruebas en India y Bangladesh y encontramos que los docentes no podían predecir con precisión los puntajes de sus estudiantes (Djaker et al., 2024). Sospechábamos que la falta de conocimiento de los docentes sobre los niveles de desempeño de sus estudiantes podría explicar por qué no dedican más atención a los estudiantes de bajo rendimiento, por lo que incluimos un conjunto ampliado de medidas en este estudio. Pedimos a cada docente que estimara el puntaje de 10 estudiantes seleccionados aleatoriamente de su aula en las evaluaciones de la intervención.¹⁶ Utilizamos estas estimaciones para calcular los ordenamientos implícitos que los docentes hacían de estos estudiantes (dado que pueden no conocer los puntajes de sus estudiantes, pero sí saber cómo se comparan entre sí) y el DE implícito dentro del aula en los puntajes de matemática (dado que pueden no saber cómo se comparan los estudiantes, pero sí intuir la variabilidad del aprendizaje en sus aulas).

Dado que estimar puntajes de pruebas es desafiante, también hicimos a los docentes preguntas más concretas, como estimar la proporción de sus estudiantes que son competentes (pueden responder correctamente la mayoría de las preguntas) en cada uno de los temas de matemática, la proporción de estudiantes que responderían correctamente ítems específicos, y si estudiantes específicos responderían correctamente a ítems específicos.

¹⁵Las Tablas C.1 y C.2 muestran la asignación de ítems.

¹⁶Antes de este ejercicio, pedimos a los docentes que indicaran qué estudiantes de una lista estaban en su aula, la cual incluía deliberadamente nombres falsos, para verificar si las estimaciones erróneas se debían a que los docentes no conocían los nombres de sus estudiantes. Durante el ejercicio, también entregamos a cada docente una copia de las evaluaciones de la intervención para reducir al mínimo la posibilidad de que las estimaciones erróneas se debieran a la falta de familiaridad con el instrumento de evaluación.

3.6. Asignación del tiempo de enseñanza e interacciones entre estudiantes y docentes

En la línea de cierre, encuestadores capacitados realizaron observaciones de aula presenciales en todas las escuelas de la muestra utilizando dos protocolos. El primero, adaptado del sistema de observación Stallings (Stallings and Mohlman, 1990; Stallings et al., 2014), codificó la proporción de cada clase de 40 minutos que el docente dedicó a la enseñanza, a la gestión del aula o a actividades ajenas a la tarea, así como los materiales que el docente estaba utilizando y los estudiantes con los cuales estaba interactuando.

El segundo, desarrollado para este estudio (Ganimian, 2023), registró *cada* interacción durante la clase—un intercambio verbal entre el docente y un estudiante o un grupo pequeño de estudiantes, en lugar de toda la clase—y consignó si fue iniciada por un docente o por un estudiante; si se formuló como afirmación o como pregunta; si fue corta (5 segundos o menos), media (entre 6 y 10 segundos) o más larga; si fue unidireccional (una persona habla), bidireccional (una persona habla y la otra responde) o multidireccional (una persona habla, la otra responde y la primera replica); si fue académica (centrada en la materia que se enseñaba), logística (centrada en aspectos de gestión de la clase) o social (centrada en la vida de los docentes o los estudiantes); si fue adversarial (el docente reprende a un estudiante), de apoyo (el docente alienta a un estudiante), mixta o ninguna de estas; y si involucró el pizarrón, los libros de texto, los cuadernos u otros materiales. Antes de cada observación, vinculamos a cada estudiante del aula con su identificador, de modo que pudiéramos comprender qué estudiantes interactúan más frecuentemente con sus docentes. Hasta donde sabemos, la nuestra es la primera medida *directa* de si los docentes de los PIBM “enseñan a la cima”. En un subconjunto de observaciones registramos también el tema de la clase, para ver si enseñar a la cima implicaba contenidos demasiado difíciles para la mayoría de los estudiantes.¹⁷

Este instrumento midió las interacciones entre estudiantes y docentes de manera confiable en dos “estudios de G(eneralizabilidad)” (Cronbach et al., 1972; Shavelson and Webb, 1991; Brennan, 2001), que mejoran los análisis tradicionales al distinguir fuentes de varianza de error (por ejemplo, ítems o evaluadores) y sus interacciones (por ejemplo, evaluadores más estrictos en algunos ítems). El primero, que incluyó a todos los docentes observados, indica que la confiabilidad de los distintos dominios puntuados (iniciador de las interacciones, dirección, contenido, tono, formulación y duración) osciló entre 0.179 y 0.481 (Tabla C.5). El segundo, para el cual agregamos un segundo evaluador en un subconjunto de aulas seleccionadas aleatoriamente, indica que la confiabilidad de los mismos dominios osciló entre 0.52 y 0.87 para los juicios relativos (determinar qué docentes interactuaron más con sus estudiantes) y entre 0.08 y 0.41 para los juicios absolutos (determinar la cantidad y los tipos específicos

¹⁷Los temas del protocolo de observación se vincularon con los marcos que utilizamos para las evaluaciones de la intervención (Tablas B.1-B.2) y de la evaluación de impacto (Tablas C.1-C.2).

de interacciones de cada docente; Tabla C.6). Una revisión reciente de estudios de G de observaciones de aula en países de ingreso alto encontró que la confiabilidad de una única observación para juicios relativos osciló entre 0 y 0.81, con una mediana de 0.38 (Ganimian et al., 2026), por lo que nuestro instrumento se compara favorablemente.

4. Estrategia empírica

Calculamos el efecto de intención de tratar (ITT, por su sigla en inglés) de que se ofrezca la intervención mediante la siguiente especificación:

$$Y_{ij} = \alpha_{r(j)} + X'_{ij}\gamma + T_j\beta + \varepsilon_{ij} \quad (1)$$

donde Y_{ij} es el resultado del estudiante i en la escuela j , $\alpha_{r(j)}$ es un efecto fijo del estrato de aleatorización, X_{ij} es un vector de covariables de línea de base seleccionadas por separado para cada resultado mediante LASSO de doble selección (Belloni et al., 2014), T_j es una variable indicadora de la asignación de la escuela j al tratamiento, y ε_{ij} es el término de error. El coeficiente β captura el efecto ITT. Estimamos la ecuación (1) por mínimos cuadrados ordinarios (MCO) con errores estándar agrupados a nivel de la escuela.

Para examinar cómo varía este efecto entre estudiantes con distinto rendimiento inicial, dividimos a los estudiantes en terciles inferior, medio y superior de la distribución del rendimiento en matemática de línea de base dentro del aula en las evaluaciones de impacto y estimamos una única regresión agrupada que incluye interacciones entre tratamiento y tercil:

$$Y_{ij} = \alpha_{r(j)} + X'_{ij}\gamma + T_j\beta_1 + \sum_{k=2}^3 \delta_k \mathbf{1}[\text{tercil}_{ij} = k] + \sum_{k=2}^3 \theta_k (T_j \times \mathbf{1}[\text{tercil}_{ij} = k]) + \varepsilon_{ij}. \quad (2)$$

El efecto de tratamiento para el tercil k se recupera como $\beta_1 + \theta_k$ (con $\theta_1 = 0$), y θ_k en sí mismo da la diferencia en los efectos de tratamiento entre el tercil k y el tercil inferior.¹⁸

¹⁸Preferimos esta especificación agrupada a las regresiones por subgrupo porque estima los coeficientes de las variables de control en la muestra completa, lo que produce errores estándar más precisos cuando los efectos de los controles son aproximadamente uniformes.

5. Resultados

5.1. Diagnóstico

5.1.1. ¿Conocen los docentes el rendimiento de sus estudiantes?

Los docentes no pueden estimar con precisión el desempeño de sus estudiantes en las pruebas. La correlación entre los puntajes de los estudiantes en las evaluaciones de la intervención de línea de base en matemática y las estimaciones de los docentes de esos puntajes fue de 0.23, lo que significa que las segundas explicaron alrededor del 5 % de la variación en los primeros (Figura 1, panel A).¹⁹ De hecho, los docentes parecen incapaces de predecir los ordenamientos de los estudiantes dentro del aula: la correlación entre los ordenamientos reales de los estudiantes y los implícitos en las estimaciones de los docentes fue de 0.16 (panel B). Por último, los DE reales dentro del aula en el rendimiento de los estudiantes y los implícitos en las estimaciones de los docentes son casi ortogonales: su correlación es de -0.01 (panel C).²⁰ Este patrón, que es consistente con nuestro estudio previo en India y Bangladesh (Djaker et al., 2024), sugiere que los docentes pueden no estar atendiendo la heterogeneidad en el rendimiento de sus estudiantes en parte porque simplemente la desconocen.

El docente típico sobreestima los puntajes de sus estudiantes de menor rendimiento. Si agrupamos a los estudiantes según su tercil en la distribución del rendimiento en matemática de línea de base dentro del aula en las evaluaciones de la intervención, el estudiante promedio del tercil inferior respondió correctamente el 37.1 % de las preguntas, pero su docente creía que había obtenido 6.9 p.p. más (panel D). Esta brecha equivale al 50 % del DE dentro del aula. En contraste, los docentes subestimaron los puntajes de los estudiantes de los terciles medio y superior en 7.9 y 19.7 p.p., respectivamente. Estos resultados—consistentes con nuestro trabajo anterior (Djaker et al., 2024) y con estudios de replicación en Benín, Nigeria y Gambia (Rodríguez-Segura, 2024), que se muestran en la Figura A.2—sugieren que los docentes pueden dedicar atención insuficiente a los estudiantes de bajo rendimiento de su aula porque no saben cuántos son ni en qué medida tienen bajo desempeño.

¹⁹La línea de base se realizó en mayo, tres meses después del inicio del año escolar (véase la sección 2.1), por lo que parece razonable esperar que los docentes estén familiarizados con el desempeño de sus estudiantes. La duración de la exposición de los docentes a los estudiantes mejora sus estimaciones, pero no demasiado: la correlación entre los puntajes de los estudiantes en las evaluaciones de impacto de cierre y las estimaciones de los docentes en el grupo de control fue de 0.33.

²⁰La dificultad de estimar los puntajes de los estudiantes no explica estos resultados. Preguntas de indagación más concretas muestran un patrón similar de precisión limitada: los docentes a menudo sobreestimaron la competencia a nivel de tema y el desempeño de estudiantes específicos en ítems específicos, aunque no en todas las preguntas a nivel de ítem (Tabla A.5).

5.1.2. ¿(Cómo) interactúan los docentes con los estudiantes?

El docente típico rara vez interactúa con sus estudiantes durante la clase. En la clase promedio de matemática de 40 minutos del grupo de control en la línea de cierre,²¹ el docente tuvo 8.5 intercambios verbales con un estudiante o un grupo pequeño de estudiantes (Figura 2, panel A). Esto sugiere que los docentes tienen pocas oportunidades de actualizar sus estimaciones del rendimiento de sus estudiantes durante la clase.

Los docentes varían ampliamente en la cantidad de interacciones con sus estudiantes. El docente con menos interacciones tuvo solo 1 y el que tuvo más registró 32; el DE entre docentes fue de 5.3 interacciones. Sin embargo, las características de los docentes explican poco de esta variación. Si agrupamos a los docentes según si poseen una característica (si es discreta) o si están por encima de la mediana (si es continua), la mayoría tiene una asociación pequeña y estadísticamente no significativa con la cantidad de interacciones, aunque los docentes de mayor edad y con más experiencia interactúan algo más frecuentemente, mientras que los docentes con títulos en educación y mayor autoeficacia interactúan menos (Tabla A.8).

La mayoría de los estudiantes no interactúa con su docente durante la clase. El estudiante de control promedio tuvo 0.24 intercambios verbales con su docente y el 79% no tuvo ninguno (Tabla A.7, panel A).²² Sin embargo, su rendimiento fue el predictor más fuerte de si interactuaron y con qué frecuencia lo hicieron (Tabla A.9). El estudiante promedio del tercil inferior de la distribución del rendimiento en matemática de línea de base dentro del aula en las evaluaciones de impacto tuvo menos de la mitad de las interacciones de su contraparte del tercil superior, y el primero tuvo 1.3 veces menos probabilidad de interactuar con su docente que el segundo (Figura 2, panel C y Tabla A.7, panel A). Así, los docentes tuvieron muy pocas oportunidades de actualizar sus estimaciones de rendimiento para este grupo.

Las diferencias en las interacciones según el rendimiento de los estudiantes se deben a que los docentes convocan menos a los estudiantes de bajo rendimiento. Los docentes de control iniciaron el 69% de las interacciones observadas, y la interacción modal consistió en un docente haciendo a un estudiante una pregunta académica y el estudiante respondiendo (Figura 2, panel B y Tabla A.7, paneles B-E). La interacción modal duró unos 10 segundos (panel F), lo que sugiere que los docentes en general hicieron preguntas de sí o no o de baja inferencia. En conjunto, estos resultados ofrecen parte de la evidencia más directa hasta la fecha de que los docentes de los PIBM enseñan a la cima de la distribución de rendimiento.

²¹No medimos las interacciones en la línea de base, por lo que usamos el grupo de control de la línea de cierre para caracterizar la enseñanza habitual en el aula durante la ronda de recolección de datos de cierre.

²²Esto es consistente con nuestras observaciones de la asignación del tiempo docente, que indican que el 69% del tiempo de clase se dedica a la enseñanza al grupo completo (Tabla A.6, panel C).

5.2. Impacto

5.2.1. ¿Funcionó la intervención como se esperaba?

Los docentes respondieron a la información de los informes como se pretendía. En las escuelas de tratamiento—las únicas que recibieron las evaluaciones de la intervención y los informes—los contenidos y los estudiantes que fueron identificados como necesitados de apoyo en los informes de la intervención de línea de base mejoraron más de la línea de base a la línea intermedia (la última vez que se administraron las evaluaciones de la intervención) que el resto. Esto es consistente con que los informes lograran enfocar la atención de los docentes.

Los estudiantes mejoraron más en los temas que los informes señalaron por bajo desempeño. Como se muestra en la Figura B.2, cada informe indicaba el tema en el que la clase tuvo el peor desempeño (por ejemplo, medición y geometría). Los estudiantes de tratamiento mejoraron más en las evaluaciones de la intervención de la línea de base a la línea intermedia en estos temas señalados que en los temas que no fueron señalados: 25 p.p. frente a 9.8 p.p.—una estimación de diferencias en diferencias (DED) de 14.5 p.p. (Tabla 1, panel A).

El mismo patrón surge para los subtemas y los ítems. Como muestra la Figura B.2, los informes identificaban los subtemas en los que los estudiantes necesitaban más ayuda (por ejemplo, aplicar propiedades geométricas para hallar un ángulo faltante). Los estudiantes de tratamiento mejoraron en estos subtemas señalados en 37.6 p.p., mientras que los no señalados mejoraron en 10.8 p.p.—un efecto DED de 19.4 p.p. (panel A). Incluso el *color* que los informes usaron para señalar los ítems parece haber tenido un efecto. Los informes mostraban en rojo los ítems con bajo desempeño (menos de un tercio de respuestas correctas), en amarillo los de desempeño medio (entre uno y dos tercios) y en verde los de desempeño alto (más de dos tercios) (Figura B.2). Los ítems rojos y amarillos mejoraron 34.7 p.p. más que los ítems verdes, y los ítems rojos mejoraron 28.1 p.p. más que los amarillos (panel A).²³

Los estudiantes señalados en los informes también mejoraron más que sus pares. La Figura B.3 muestra que los informes identificaban a los estudiantes con mayor necesidad de apoyo en cada tema. Los estudiantes de tratamiento señalados en al menos un tema mejoraron 22.5 p.p. de la línea de base a la línea intermedia, en comparación con 12.7 p.p. entre los estudiantes no señalados en ningún tema—una estimación DED de 9.9 p.p. (Tabla 1, panel B). Las mejoras fueron mayores para los estudiantes señalados en más temas: los señalados en uno, dos y tres temas mejoraron 21.2, 25.5 y 33.3 p.p., respectivamente. Las estimaciones DED correspondientes fueron de 8.5, 12.8 y 20.6 p.p., y la última se basa en solo 27 estudiantes.

²³Dado que los colores de los ítems se asignaron mecánicamente a partir del desempeño de la clase en la línea de base, los ítems rojos y amarillos comenzaron con puntajes más bajos que los verdes y por lo tanto tenían más margen para mejorar. Estas comparaciones pueden reflejar en parte, entonces, una reversión hacia la media o efectos de techo, más que a una respuesta de los docentes a los colores.

Además, los estudiantes señalados mejoraron más en los temas específicos por los que fueron señalados. Entre los estudiantes de tratamiento señalados en al menos un tema, el desempeño aumentó 41.8 p.p. en los temas señalados, en comparación con 9.8 p.p. en los temas en los que no fueron señalados—una estimación DED de 32.9 p.p. (panel D, filas de temas señalados). Dado que el señalamiento reflejaba el desempeño de línea de base, esta comparación sigue siendo descriptiva. Aun así, el patrón es consistente con que los informes enfocaran la atención docente en los temas y estudiantes señalados para recibir apoyo.

5.2.2. ¿Elevó la intervención el rendimiento en matemática?

La intervención tuvo efectos amplios sobre el rendimiento en matemática. El estudiante promedio mejoró 0.23 DE en matemática y no podemos descartar que los estudiantes del tercio inferior de la distribución del rendimiento en matemática de línea de base dentro del aula en las evaluaciones de impacto hayan mejorado tanto como los de los otros dos tercios (Tabla 2, panel A). Los gráficos no paramétricos de los efectos promedio de tratamiento confirman este patrón, con los estudiantes de tratamiento superando a sus pares de control por un margen similar en toda la distribución de línea de base (Figura A.4, panel A). Los efectos de tratamiento por cuantiles presentan un patrón similar, pero con un contraste menor (panel B).

La intervención mejoró el rendimiento en todos los temas y habilidades cognitivas evaluados. Los efectos fueron de 0.24 DE en número, 0.16 DE en medición y geometría, y 0.16 DE en datos (Tabla 2, panel B). Fueron de 0.18 DE en conocer, 0.19 DE en aplicar y 0.26 DE en razonar (panel C). Aunque las estimaciones puntuales fueron mayores en número y razonar, no podemos rechazar que los efectos fueran iguales entre temas ($p = 0.131$) o entre habilidades ($p = 0.144$); las mejoras fueron, por lo tanto, de amplio alcance.

De manera consistente con la sección 5.2.1, los estudiantes señalados tuvieron mejores resultados que sus contrapartes de control. Comparar los señalamientos observados en el tratamiento con los señalamientos de las evaluaciones de impacto entre los controles, o aplicar los señalamientos de las evaluaciones de impacto en ambos grupos, arroja efectos similares: 0.21 y 0.24 DE, respectivamente (Tabla A.11). Los estudiantes de tratamiento no señalados también superaron a sus contrapartes de control por márgenes similares; no podemos rechazar la igualdad bajo ninguna de las dos especificaciones. Así, todo lo que haya beneficiado a los estudiantes focalizados también benefició a sus pares que no fueron focalizados.

5.2.3. ¿Tuvo la intervención efectos de derrame sobre la lectura?

La intervención también tuvo efectos de derrame positivos sobre la lectura. El estudiante promedio mejoró 0.17 DE en lectura (Tabla 3, panel A). Los efectos son estadísticamente significativos únicamente para los terciles medio y superior de la distribución del rendimiento en matemática de línea de base dentro del aula en las evaluaciones de impacto, pero no

podemos rechazar las hipótesis nulas de que los estudiantes del tercil inferior hayan mejorado en la misma medida que sus contrapartes de los terciles medio y superior.

Las mejoras se concentraron en los insumos y las habilidades que ejercitan los problemas con enunciado de matemática. Los docentes de 6.º grado en Bangladesh son especializados por materia, por lo que es poco probable que la intervención haya afectado el desempeño en lectura cambiando la forma en que se enseñaba lectura. Por lo tanto, verificamos si mejoró el rendimiento en los temas y las habilidades que podrían trasladarse entre materias. Eso es lo que encontramos: los efectos fueron de 0.15 DE en los ítems de palabras y 0.13 DE en los ítems de oraciones (panel B) y de 0.15 DE en vocabulario y 0.14 DE en los ítems de comprensión (panel C). Estos resultados sugieren que, a través de cambios en las prácticas de enseñanza de los docentes de matemática (por ejemplo, un mayor uso de problemas con enunciado), los estudiantes aprendieron insumos y habilidades que mejoraron su competencia lectora.²⁴

5.2.4. ¿Elevó la intervención el esfuerzo y la motivación de los estudiantes?

Encontramos alguna evidencia de que la intervención aumentó el esfuerzo y la motivación de los estudiantes, particularmente entre los estudiantes de menor rendimiento. Los registros escolares indicaron que los estudiantes asistieron el 70 % de los días tanto en la línea intermedia como en la de cierre; los encuestadores encontraron presentes al 62 % de los matriculados el día de la encuesta en la línea intermedia y al 61 % en la de cierre. La intervención no afectó en promedio ninguna de las dos medidas en ninguna ronda; de hecho, podemos descartar aumentos superiores a 4 p.p. en la asistencia basada en registros y a 5.0 p.p. en la asistencia observada. En la línea de cierre, sin embargo, aumentó la asistencia basada en registros escolares entre los estudiantes de bajo rendimiento en 3.5 p.p. (Tabla A.15).²⁵

La intervención también mejoró las habilidades socioemocionales. Aumentó la autoeficacia (confianza en completar tareas) en 0.14 DE, la mentalidad de crecimiento (creencia de que la habilidad y la inteligencia pueden desarrollarse) en 0.11 DE, y la orientación a metas de tarea (deseo de aprender y dominar el material de la clase) en 0.10 DE (Tabla 4). Los efectos fueron nulos para la autogestión (preparación, atención y autocontrol), el miedo al fracaso (preocupación y dudas sobre sí mismo al fracasar), la motivación por dominar tareas (trabajo duro y persistencia) y la determinación (perseverancia de largo plazo), aunque no podemos descartar efectos positivos de 0.10 DE para ninguna de ellas. Entre los estudiantes de bajo rendimiento, los aumentos respectivos fueron de 0.14, 0.20 y 0.13 DE; solo la mentalidad de crecimiento difirió significativamente de los estudiantes de alto rendimiento.

²⁴Las mejoras en lectura también fueron similares para los estudiantes señalados para recibir apoyo en matemática y los no señalados bajo ambos enfoques utilizados para definir grupos señalados comparables (Tabla A.12). Dado que las evaluaciones de la intervención cubrieron solo matemática, estas comparaciones utilizan los señalamientos basados en matemática que se definen en la sección 3.4.

²⁵También verificamos si la intervención afectó el esfuerzo dedicado a las evaluaciones de cierre, pero hallamos resultados nulos, porque casi todos los estudiantes intentaron responder casi todos los ítems.

5.3. Mecanismos

5.3.1. ¿Actualizaron los docentes sus estimaciones del rendimiento de los estudiantes?

Las estimaciones de los docentes no se volvieron más precisas. La intervención no impactó las brechas absolutas entre los puntajes u ordenamientos estimados y los reales (Tabla 5, panel A), ni la brecha entre los DE estimados y los reales de los puntajes dentro del aula (panel C). Esto quizás no sorprenda porque los informes identificaban los temas y los estudiantes con mayor necesidad de apoyo. No revelaban los puntajes individuales ni la distribución dentro del aula necesaria para actualizar las creencias de los docentes sobre el desempeño absoluto de los estudiantes o su posición relativa en su clase en la indagación de la línea de cierre.

En cambio, la intervención desplazó a la baja las estimaciones de los docentes a lo largo de la distribución de rendimiento. Los puntajes estimados se desplazaron a la baja respecto de los puntajes reales en 0.47 DE dentro del aula, mientras que los rangos percentilares estimados dentro del aula se desplazaron a la baja respecto de los rangos reales en 0.9 p.p. (Tabla 5, panel B). El desplazamiento de los puntajes fue similar entre los terciles de rendimiento. Junto con los impactos positivos en los tres terciles de rendimiento, y entre los estudiantes señalados y no señalados en las evaluaciones de impacto, la ausencia de una actualización diferencial sugiere que lo que hayan hecho los docentes para mejorar el desempeño de los estudiantes de bajo rendimiento también benefició a sus pares de mayor rendimiento.

5.3.2. ¿Cambió la intervención las interacciones entre estudiantes y docentes?

Consistentemente con esos desplazamientos a la baja en todos los terciles, la intervención los llevó a distribuir las interacciones de manera más uniforme entre los estudiantes. Dejó sin cambios el promedio de interacciones por estudiante por clase (Tabla 6, panel A), pero las interacciones aumentaron en 0.01 para los estudiantes de bajo rendimiento y cayeron en 0.02 y 0.07 para los de rendimiento medio y alto, con una diferencia significativa entre bajo y alto. Estos cambios implican que la proporción de las interacciones que recibieron los estudiantes de bajo rendimiento aumentó aproximadamente tanto (del 22 % al 27 %) como cayó la de los estudiantes de alto rendimiento (del 45 % al 40 %).

5.3.3. ¿Cambió la intervención los temas que cubrieron los docentes?

La intervención parece haber llevado a los docentes a centrarse en contenidos más fundamentales. En una submuestra aleatoria, los observadores registraron si los docentes cubrían número, medición y geometría, o datos. La intervención redujo la cobertura de datos en 15.3 p.p. desde el 18.2 %, dejó sin cambios medición y geometría, y aumentó la cobertura de número en 13.6 p.p. desde el 72.7 % (Tabla A.14). Este último efecto no fue

estadísticamente significativo, pero el pequeño tamaño de la submuestra de observaciones implica que la estimación es imprecisa y resulta difícil de distinguir de cero.

En conjunto, los tres mecanismos sugieren que la intervención cambió cómo los docentes enseñaban a toda la clase, más que inducir sólo apoyo adicional para los estudiantes de bajo rendimiento. Los docentes revisaron sus estimaciones a la baja en toda la distribución, distribuyeron las interacciones de manera más uniforme y se centraron en contenidos más fundamentales. Esta interpretación es consistente con los impactos positivos en todos los terciles de rendimiento y tanto en los estudiantes señalados como en los no señalados.

6. Discusión

6.1. Costo-efectividad

La intervención de evaluación formativa fue altamente costo-efectiva. Siguiendo el método de ingredientes de Dhaliwal et al. (2013), que identifica y valora cada insumo requerido para implementar la intervención, estimamos que dos rondas de evaluación, elaboración de informes y orientación docente costaron USD 5.06 por estudiante.²⁶ Esto implica una estimación de costo-efectividad de 4.54 DE en matemática y 3.36 DE en lectura por cada USD 100 gastados. Estas estimaciones están en los percentiles 86 y 80 de la distribución de estimaciones de costo-efectividad de intervenciones educativas rigurosamente evaluadas en los PIBM (Angrist et al., 2025). Las mejoras siguen siendo grandes en años de escolaridad ajustados por aprendizaje (LAYS), tomando 0.8 DE de aprendizaje como referencia de un año de escolarización de alta calidad (Angrist et al., 2020). Los efectos equivalen a 5.68 LAYS en matemática y 4.20 en lectura por USD 100—más de 11 y ocho veces la mediana de referencia de 0.49 LAYS.

Ninguna base de datos exhaustiva registra la costo-efectividad de las intervenciones socioemocionales en los PIBM, por lo que comparamos nuestros resultados con otros de la economía. La intervención generó 2.67 DE en autoeficacia, 2.17 en mentalidad de crecimiento y 1.98 en orientación a metas de tarea por cada USD 100 gastados. Estas estimaciones superan los 1.36-1.59 DE de las visitas domiciliarias a través del sistema de nutrición de Bangladesh (Bos et al., 2024) y son similares a los 3.29-3.41 DE de un currículo de toma de perspectiva implementado con estudiantes de primaria en Turquía (Alan et al., 2021).²⁷

²⁶El costo total fue de 1.87 millones de BDT, o USD 18,155 en dólares de 2023, para aproximadamente 3,588 estudiantes. Los costos excluyen las actividades de evaluación de impacto, los costos de oportunidad de los docentes y los insumos existentes del sistema escolar propios de la escolarización habitual.

²⁷Otros estudios han encontrado impactos sobre las habilidades socioemocionales (Alan and Ertac, 2018; Alan et al., 2019; Özler et al., 2018; Premand and Barry, 2022), pero no reportan los costos de los programas.

6.2. Escalamiento

Es fiscalmente factible que Bangladesh escale la intervención de evaluación formativa a nivel nacional. En 2023, tenía 2.10 millones de estudiantes de 6.º grado matriculados en escuelas secundarias generales (BANBEIS, 2023). Si multiplicamos esta cifra por nuestro costo por estudiante de 521 BDT, extender la intervención a todos los estudiantes de sexto grado costaría 1.10 mil millones de BDT (USD 10.6 millones), lo que representa el 0.23 % de la asignación del gobierno para educación secundaria y superior del año escolar 2025-2026 (Ministry of Finance, 2025). Sin embargo, esta proyección sobreestima los costos recurrentes dadas las economías de escala que generarían la adquisición y los informes centralizados.

Un escalamiento a nivel nacional podría aprovechar la estructura administrativa de Bangladesh, pero puede requerir una fuerza de trabajo de campo temporal considerable. En nuestro estudio, 26 encuestadores y 6 supervisores implementaron dos rondas de evaluación en 156 escuelas de tratamiento, con cada persona desplegada 17 días. Aplicando estas razones de personal a las 20,448 instituciones de secundaria general del país se obtienen aproximadamente 3,400 encuestadores y 800 supervisores, o 71,000 días-persona por año (BANBEIS, 2023). Las 480 oficinas de educación secundaria de *upazila* del país, junto con sus oficinas distritales y zonales, proveen una estructura existente para coordinar este esfuerzo; cada oficina de *upazila* supervisaría 43 escuelas en promedio (DSHE, 2026). Sin embargo, no sabemos si estas oficinas tienen personal suficiente para administrar y supervisar las evaluaciones. Escalar el modelo evaluado puede requerir, por lo tanto, contratar evaluadores externos.

Bangladesh está comenzando a incorporar la evaluación formativa y la remediación en su sistema de educación secundaria, pero esta iniciativa difiere de la intervención que evaluamos. El programa apoyado por el Banco Mundial, aprobado en diciembre de 2025, busca capacitar a los docentes de 6.º y 8.º grado para administrar evaluaciones formativas en matemática y bangla y usar los resultados de las pruebas para implementar educación remedial. Prevé alcanzar 1,200 escuelas en su primer año (para marzo de 2026), 3,000 escuelas en su segundo año y 7,200 escuelas en su tercer año (World Bank, 2023, 2025). Esperamos que esta implementación arroje luz sobre si pedir a los docentes que evalúen a los estudiantes, analicen los resultados y ajusten su enseñanza puede tener resultados similares al enfoque que evaluamos. Investigaciones previas de India señalan dos desafíos principales: asegurar que las pruebas administradas por los docentes sean tan confiables como las realizadas por evaluadores externos (Singh and Berg, 2024) y que los docentes usen los resultados de las pruebas para informar sus prácticas de enseñanza (Duflo et al., 2015).

6.3. Información diagnóstica

Nuestro estudio avanza de manera significativa la evidencia global sobre el impacto de la información diagnóstica (de bajas consecuencias) en los PIBM.²⁸ Aprovechó elementos comunes de intervenciones previas en Argentina (de Hoyos et al., 2021; de Hoyos et al., 2023) e India (Banerjee et al., 2018) y mostró que las intervenciones que son breves (los informes tenían cinco páginas), fáciles de usar (los resultados se presentaban en palabras, sorteando los desafíos que los docentes pueden enfrentar al interpretar gráficos), concretas (los informes destacaban los temas y los estudiantes con mayor necesidad de apoyo) y accionables (las claves de respuestas explicaban tanto las respuestas correctas como las incorrectas) pueden elevar las habilidades académicas y socioemocionales. Verificó que los docentes respondieron a la información específica incluida en las intervenciones, siguiendo cuánto mejoraron los estudiantes en los aspectos destacados en los informes en relación con otros. E identificó tres mecanismos de impacto: la conciencia de los docentes sobre el aprendizaje de sus estudiantes, sus interacciones con los estudiantes y el nivel al que calibran la enseñanza en el aula.

Notablemente, nuestro estudio demuestra que la provisión de información puede impactar la enseñanza y el aprendizaje incluso en ausencia de cambios en los incentivos institucionales que enfrentan los docentes. Junto con la evidencia alentadora de intervenciones similares (de Hoyos et al., 2021; de Hoyos et al., 2023), sugiere que los efectos nulos anteriores (Muralidharan and Sundararaman, 2010) no pueden atribuirse únicamente a la falta de incentivos vinculados a la información, como se creía previamente. Además, es menos disruptiva política y administrativamente que los esfuerzos por cambiar los incentivos, como las reformas curriculares (Rodriguez-Segura and Mbiti, 2022), el agrupamiento por nivel (Duflo et al., 2011), el pago por mérito (Muralidharan and Sundararaman, 2011) o los docentes contratados (Duflo et al., 2015). De hecho, la mayoría de los PIBM ya recolecta los datos necesarios para las intervenciones de información diagnóstica. Para 2024, 103 de 131 de esos países habían reportado administrar una evaluación nacionalmente representativa de lectura y matemática al final de la primaria o la secundaria (UNESCO, 2026).

6.4. Enseñar a la cima

Hasta donde sabemos, nuestro estudio también aporta la primera evidencia directa de que los docentes de los PIBM “enseñan a la cima” de la distribución de habilidades. Documenta que los docentes no solo interactúan con los estudiantes de alto rendimiento más a menudo que con los de bajo rendimiento, sino que esta distribución desigual de las interacciones

²⁸Consideramos que estos esfuerzos, en los que se brindan a los directores o docentes datos granulares y accionables (por ejemplo, resultados a nivel de tema, habilidad o ítem) para guiar su práctica, prueban una teoría del cambio distinta de la de aquellos que divulgan los puntajes de escuelas o estudiantes (Banerjee et al., 2010; Mizala and Urquiola, 2013; Andrabi et al., 2017; Camargo et al., 2018) para promover la rendición de cuentas a través de la demanda de los padres y de la competencia entre las escuelas.

se explica por las interacciones que inician los docentes (no por las de los estudiantes). La evidencia más fuerte de este fenómeno provenía hasta ahora de un experimento en Kenia que encontró patrones de rendimiento consistentes con docentes calibrando la enseñanza a la cima de clases heterogéneas, pero que no observó sus decisiones de enseñanza (Duflo et al., 2011).²⁹ Complementamos esta evidencia midiendo las creencias y prácticas de aula de los docentes.

También mostramos que la tendencia a enseñar a la cima es más maleable de lo que muchos pensaban. La intervención que evaluamos impactó tanto la forma en que los docentes asignaban las interacciones entre estudiantes de bajo, medio y alto rendimiento como la (dificultad de los) temas que enseñaban. Esto es llamativo, dado que estos comportamientos se han atribuido a que los docentes buscan evitar que se les exija rendir cuentas por no completar el currículo o por que muy pocos de sus estudiantes aprueben los exámenes de altas consecuencias (Glewwe et al., 2009; Kremer et al., 2009; Pritchett and Beatty, 2015). Nuestros resultados deberían ampliar nuestra consideración de los incentivos que enfrentan los docentes. Por ejemplo, es del todo plausible que los docentes de nuestro contexto vean la intervención como un cambio en las prioridades del gobierno y estén respondiendo a ese cambio percibido.

6.5. Enseñanza diferenciada

Finalmente, nuestros resultados plantean la posibilidad de que una proporción no trivial del efecto de la enseñanza diferenciada pueda deberse a que los docentes actualizan sus creencias previas sobre los niveles de aprendizaje de sus estudiantes. Según nuestra interpretación, la diferenciación implica evaluar a los estudiantes, agruparlos según su desempeño en las pruebas y asignar a cada grupo actividades que se correspondan con su nivel de preparación. Junto con otros enfoques que buscan mejorar la adecuación de la enseñanza a las necesidades de los estudiantes—conocidos colectivamente como “Enseñar al Nivel Adecuado” (TaRL)—se considera una de las intervenciones educativas más efectivas del mundo (Akyeampong et al., 2023; Angrist and Meager, 2023). En nuestro estudio, el primero de estos pasos (evaluar a los estudiantes y reportar los resultados de las pruebas a los docentes) llevó a los docentes a ajustar a la baja sus estimaciones de los puntajes de los estudiantes, a interactuar con los estudiantes de manera más equitativa y a enseñar temas más fundamentales, elevando el aprendizaje de todos los estudiantes. Estos resultados nos invitan a considerar la posibilidad de que la conciencia de los docentes sobre el rendimiento de sus estudiantes pueda dar cuenta de una proporción no trivial del impacto de la enseñanza diferenciada.

²⁹Los autores sostienen que dos patrones son inconsistentes con que los docentes calibren la enseñanza al estudiante mediano y consistentes con que se centren en los de la cima: el agrupamiento por nivel elevó el rendimiento a lo largo de la distribución de línea de base, incluso en el grupo inferior; y los estudiantes cercanos a la mediana tuvieron un desempeño similar en ambos grupos a pesar de las grandes diferencias en el rendimiento de sus pares. También encuentran mejoras mayores en habilidades básicas entre los estudiantes de menor rendimiento y en habilidades avanzadas entre los de mayor rendimiento, lo que es consistente con que los docentes ajusten la enseñanza a cada grupo.

La evidencia creciente sugiere que los docentes de los PIBM desconocen el rendimiento de sus estudiantes. En un trabajo previo en India y Bangladesh (Djaker et al., 2024), fuimos de los primeros en mostrar que los docentes de los PIBM tienen una probabilidad mucho menor de estimar correctamente el rendimiento de sus estudiantes que sus pares de los países de ingreso alto (Hoge and Coladarci, 1989; Südkamp et al., 2012; Kaufmann, 2024).³⁰ Ese estudio dio lugar a replicaciones en Benín, Nigeria y Gambia (Rodríguez-Segura, 2024), que demostraron que los patrones que observamos en el sur de Asia se extendían a África subsahariana. En el presente estudio, mostramos que estas estimaciones responden a la información sobre el rendimiento de los estudiantes y que el ajuste uniforme a la baja en las estimaciones coincidió con una asignación más igualitaria de las interacciones entre estudiantes y docentes y con una mayor cobertura de temas fundamentales. Este patrón se hace eco del de un estudio reciente en India que encontró que los docentes sobreestimaban el rendimiento de sus estudiantes, pero que después de implementar un programa de educación remedial—que implicaba evaluar a sus estudiantes—ajustaron sus estimaciones a la baja (Beg et al., 2024). Estos dos estudios plantean si los efectos de la diferenciación pueden deberse a que los docentes actualizan sus creencias previas sobre las habilidades estudiantiles y ajustan la enseñanza en consecuencia.

Esta pregunta es crucial porque las evaluaciones de impacto previas de la enseñanza diferenciada han encontrado que los docentes tienen dificultades para implementar uno de sus componentes clave (agrupar a los estudiantes según su desempeño en las pruebas) durante la jornada escolar a menos que se les brinde tiempo dedicado o adicional. En Bihar, India, solo el 1.6 % de los docentes agrupó a los estudiantes como se pretendía cuando se les asignaron materiales y el 3.8 % lo hizo cuando además recibieron capacitación. En Uttarakhand, solo el 10 % de los docentes con capacitación y materiales agrupó a los estudiantes y el 5.1 % lo hizo cuando además se les otorgó apoyo de voluntarios. En contraste, en Haryana el 91 % de los docentes agrupó a los estudiantes cuando se les dio una hora dedicada a la diferenciación, en comparación con apenas el 2 % de ellos fuera de esa hora (Banerjee et al., 2017).³¹ Dados los impactos de evaluar a los estudiantes y reportar los resultados a los docentes, y los desafíos de agrupar a los estudiantes según su desempeño en las pruebas, nuestro estudio sugiere que una versión simplificada de la diferenciación podría aumentar la adopción por parte de los docentes, la costo-efectividad y el potencial de escalamiento.

³⁰Wadmare et al. (2022) habían encontrado que el 40 % de los docentes creía erróneamente que sus estudiantes de bajo rendimiento habían alcanzado la alfabetización fundamental. Sin embargo, su muestra impedía comparar los errores entre grupos de rendimiento, su foco en las habilidades fundamentales pudo haber subestimado las brechas respecto de las expectativas curriculares, y su uso de niveles de desempeño produjo una medida gruesa de la correspondencia entre el rendimiento real y el estimado.

³¹Incluso el tiempo dedicado puede no asegurar la implementación si los docentes lo ven como algo que compite con sus otras responsabilidades. En Ghana, solo el 6 % de los docentes agrupó a los estudiantes aun cuando se suponía que debía dedicarse una hora a ello. Los docentes reportaron que todavía se esperaba que cubrieran el currículo del nivel del grado (Duflo et al., 2024).

7. Conclusión

Los docentes no pueden alinear la enseñanza con las necesidades de los estudiantes cuando carecen de información accionable sobre lo que sus clases comprenden. Aportamos evidencia experimental de que dos rondas de evaluaciones formativas, informes y orientación pedagógica ayudaron a los docentes de Bangladesh a responder a esta restricción. La intervención elevó el rendimiento en matemática en 0.23 DE y en lectura en 0.17 DE, mejoró varias habilidades socioemocionales y generó mejoras similares a lo largo de la distribución de rendimiento de línea de base, incluso entre los estudiantes que no fueron identificados como necesitados de apoyo en los informes y otros compañeros de las mismas aulas a lo largo del año escolar.

La intervención parece haber funcionado reorientando la atención de los docentes. Los docentes ajustaron a la baja sus estimaciones del desempeño de sus estudiantes de manera generalizada, lo que los llevó a distribuir las interacciones de manera más equitativa y a repasar contenidos fundamentales. Las mejoras amplias indican que centrarse en las necesidades de aprendizaje no atendidas puede ayudar a los estudiantes de menor rendimiento sin reducir las mejoras de sus compañeros en otros puntos de la distribución de rendimiento inicial.

En conjunto, la información de las evaluaciones utilizable para la enseñanza, combinada con orientación concreta, ofrece una vía promisorio para ayudar a los docentes a atender aulas académicamente diversas. A un costo de USD 5.06 por estudiante, la implementación a nivel nacional parece fiscalmente factible, pero replicar el modelo evaluado requeriría una evaluación independiente confiable y una capacidad de implementación sustancial. Investigaciones futuras deberían probar si los sistemas escolares pueden preservar ambas en la implementación rutinaria a escala sin sacrificar la confiabilidad de las evaluaciones ni la utilidad pedagógica de los informes para los docentes a lo largo del tiempo.

Referencias

- Abadzi, H. (2009). Instructional time loss in developing countries: Concepts, measurement, and implications. *The World Bank Research Observer* 24(2), 267–290.
- Aghion, P. and J. Tirole (1997). Formal and real authority in organizations. *Journal of Political Economy* 105(1), 1–29.
- Akyeampong, T., T. Andrabi, A. V. Banerjee, R. Banerji, S. Dynarski, R. Glennerster, S. Grantham-McGregor, K. Muralidharan, B. Piper, S. Ruto, J. Saavedra, S. Schmelkes, and H. Yoshikawa (2023). *2023 cost-effective approaches to improve global learning. What does recent evidence tell us are “smart buys” for improving learning in low- and middle-income countries?* London, UK; Washington, DC; New York, NY: Foreign, Commonwealth & Development Office (FCDO), World Bank, United Nations International Children’s Emergency Fund (UNICEF), United States Agency for International Development (USAID). London, UK; Washington, DC; New York, NY: Foreign, Commonwealth & Development Office (FCDO), World Bank, United Nations International Children’s Emergency Fund (UNICEF), United States Agency for International Development (USAID).
- Alan, S., C. Baysan, M. Gumren, and E. Kubilay (2021). Building social cohesion in ethnically mixed schools: An intervention on perspective taking. *The Quarterly Journal of Economics* 136(4), 2147–2194.
- Alan, S., T. Boneva, and S. Ertac (2019). Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. *The Quarterly Journal of Economics* 134(3), 1121–1162.
- Alan, S. and S. Ertac (2018). Fostering patience in the classroom: Results from randomized educational intervention. *Journal of Political Economy* 126(5), 1865–1911.
- Andrabi, T., J. Baron, I. Macdonald, and Z. Qureshi (2025). Free to choose: Introducing technology for foundational learning in Pakistan. Working paper, April 13.
- Andrabi, T., J. Das, and A. I. Khwaja (2017). Report cards: The impact of providing school and child test scores on educational markets. *American Economic Review* 107(6), 1535–1563.
- Angrist, N., C. Cullen, and J. Magat (2025). Cheaper (and more effective) by the dozen: Evidence from 12 randomized a/b tests optimizing tutoring for scale. edworkingpaper no. 25-1333. *Annenberg Institute for School Reform at Brown University*.
- Angrist, N., S. Djankov, P. K. Goldberg, and H. A. Patrinos (2021). Measuring human capital using global learning data. *Nature* 592(7854), 403–408.
- Angrist, N., D. K. Evans, D. Filmer, R. Glennerster, F. H. Rogers, and S. Sabarwal (2020). How to improve education outcomes most efficiently? A comparison of 150 interventions using the new learning-adjusted years of schooling metric. (Policy Research Working Paper No. 9450). Washington, DC: The World Bank.

- Angrist, N., D. K. Evans, D. Filmer, R. Glennerster, H. Rogers, and S. Sabarwal (2025). How to improve education outcomes most efficiently? a review of the evidence using a unified metric. *Journal of Development Economics* 172, 103382.
- Angrist, N. and R. Meager (2023). Implementation matters: Generalizing treatment effects in education.
- BANBEIS (2023). Bangladesh education statistics 2023. Dhaka, Bangladesh: Bangladesh Bureau of Educational Information and Statistics (BANBEIS). Retrieved from <https://banbeis.portal.gov.bd/sites/default/files/files/banbeis.portal.gov.bd/npfblock/Bangladesh%20Education%20Statistics%202023%20%281%29.pdf>.
- Banerjee, A. V., R. Banerji, J. Berry, E. Duflo, H. Kannan, S. Mukerji, M. Shotland, and M. Walton (2017). From proof to concept to scalable policies: Challenges and solutions, with an application. *Journal of Economic Perspectives* 31(4), 73–102.
- Banerjee, A. V., R. Banerji, E. Duflo, R. Glennerster, and S. Khemani (2010). Pitfalls of participatory programs: Evidence from a randomized evaluation in education in india. *American Economic Journal: Economic Policy* 2(1-30), 1–30.
- Banerjee, A. V., S. Cole, E. Duflo, and L. L. Linden (2007). Remedying education: Evidence from two randomized experiments in india. *The Quarterly Journal of Economics* 122(3), 1235–1264.
- Banerjee, A. V. and E. Duflo (2011). Top of the class. *Poor economics: A radical rethinking of the way to fight global poverty*. Philadelphia, PA: Public Affairs.
- Banerjee, A. V., A. J. Ganimian, and S. Mukerji (2018). The impact of diagnostic feedback on differentiated instruction: Experimental evidence from India. *Unpublished manuscript*. New Delhi, India: Abdul Latif Jameel Poverty Action Lab (J-PAL) South Asia.
- Barro, R. J. and J. W. Lee (2026). Barro-Lee educational attainment indicators. World Bank Education Statistics, Barro-Lee educational attainment indicators. Retrieved from <https://databank.worldbank.org/source/education-statistics-%5E-all-indicators>. Retrieved: May 21, 2026.
- Beg, S. A., A. E. Fitzpatrick, J. T. Kerwin, A. Lucas, and K. W. Rahman (2024, December). When given discretion teachers did not shirk: Evidence from remedial education in secondary schools. Working Paper 33242, National Bureau of Economic Research.
- Beg, S. A., A. E. Fitzpatrick, and A. M. Lucas (2024). Managing to learn.
- Beg, S. A., A. M. Lucas, W. Halim, and U. Saif (2022). Engaging teachers with technology increased achievement, bypassing teachers did not. *American Economic Journal: Economic Policy* 14(2), 61–90.
- Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies* 81(2), 608–650.
- Bergman, P. (2020). Parent-child information frictions and human capital investment: Evidence from a field experiment. *Journal of Political Economy*.

- Bhattacharjea, S., W. Wadhwa, and R. Banerji (2011). Inside primary schools: A study of teaching and learning in rural india.
- Bold, T., D. Filmer, G. Martin, E. Molina, B. Stacy, C. Rockmore, J. Svensson, and W. Wane (2017). Enrollment without learning: Teacher effort, knowledge, and skill in primary schools in africa. *Journal of Economic Perspectives* 31(4), 185–204.
- Bos, J. M., A. S. Shonchoy, S. Ravindran, and A. Khan (2024). Early childhood human capital formation at scale. *Journal of Public Economics* 231, 105046.
- BRAC (2015). Education watch 2014: Whither grade V examination? An assessment of primary education completion examination in Bangladesh. Dhaka, Bangladesh: Building Resources Across Communities (BRAC).
- Brennan, R. L. (2001). *Generalizability Theory*. New York, NY: New York: Springer-Verlag.
- Bruns, B. and J. Luque (2014). *Great teachers: How to raise student learning in Latin America and the Caribbean*. Washington, DC: Washington, DC: The World Bank.
- Büchel, K., M. Jakob, C. Kühnhanss, D. Steffen, and A. Brunetti (2020). The relative effectiveness of teachers and learning software: Evidence from a field experiment in El Salvador. (UBSS Working Paper No. 36). Bern, Switzerland: University of Bern.
- Camargo, B., R. Camelo, S. Firpo, and V. Ponczek (2018). Information, market incentives, and student performance: Evidence from a regression discontinuity design in Brazil. *The Journal of Human Resources* 53(2), 414–444.
- Cilliers, J. and J. Habyarimana (2026). Addressing weak links in government implementation at scale: Experimental evidence from a school governance reform in Tanzania. *Journal of Development Economics* 182, 103824.
- Cronbach, L. J., G. C. Gleser, H. Nanda, and N. Rajaratnam (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.
- de Barros, A. and T. Lubozha (2026). Targeting foundational skills at scale: Skill specificity and transfer. CEGA Working Paper No. 259. Berkeley, CA: Center for Effective Global Action (CEGA).
- de Hoyos, R., S. Djaker, A. J. Ganimian, and P. A. Holland (2023). Can performance-management tools and training complement diagnostic feedback? experimental evidence from public schools in argentina. *Economics of Education Review* 9, 102518.
- de Hoyos, R., A. J. Ganimian, and P. A. Holland (2021). Teaching *with* the test: Experimental evidence on diagnostic feedback and capacity-building for schools in Argentina. *World Bank Economic Research* 35(2), 499–520.
- Dhaliwal, I., E. Duflo, R. Glennerster, and C. Tulloch (2013). Comparative cost-effectiveness analysis to inform policy in developing countries: A general framework with applications for education. *Education Policy in Developing Countries*, 285–338.

- Dizon-Ross, R. (2019). Parents’ beliefs and children’s education: Experimental evidence from Malawi. *American Economic Review* 109(8), 2728–2765.
- Djaker, S. (2023). *Targeting teachers’ perceptions of student ability to improve learning levels in low- and middle-income countries*. Ph. D. thesis, New York University.
- Djaker, S., A. J. Ganimian, and S. Sabarwal (2024). Out of sight, out of mind? the gap between students’ test performance and teachers’ estimations in india and bangladesh. *Economics of Education Review* 102, 102575.
- DSHE (2019). Frameworks for National Assessment of Secondary Students 2019. Subject: Bangla, English, and Mathematics. Grades: 6, 8, and 10. Dhaka, Bangladesh: Monitoring and Evaluation Wing, Directorate of Secondary and Higher Education (DSHE).
- DSHE (2026). History and administrative structure. Dhaka, Bangladesh: Directorate of Secondary and Higher Education (DSHE). Retrieved from <https://dshe.gov.bd/pages/static-pages/691997c4933eb65569ddef95>. Retrieved: August 4, 2026.
- Duflo, A., J. Kiessel, and A. M. Lucas (2024). Experimental evidence on alternative policies to increase learning at scale. *The Economic Journal* 134(661), 1985–2008.
- Duflo, E., J. Berry, S. Mukerji, and M. Shotland (2015). A wide angle view of learning: Evaluation of the CCE and LEP programmes in Haryana, India. (Impact Evaluation Report No. 22). International Initiative for Impact Evaluation (3ie). New Delhi, India.
- Duflo, E., P. Dupas, and M. Kremer (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya. *American Economic Review* 101(5), 1739–1774.
- Duflo, E., P. Dupas, and M. Kremer (2015). School governance, teacher incentives, and pupil-teacher ratios: Experimental evidence from kenyan primary schools. *Journal of Public Economics* 123, 92–110.
- Evans, D. K. and A. M. Acosta (2026). Changing perceptions of educational returns in low-and middle-income countries: A review of the evidence. *International Journal of Educational Development* 122, 103563.
- Ganimian, A. J. (2023). Do teachers in developing countries “teach to the top”? A new approach to measure student-teacher interactions. Spencer Foundation/National Academy of Education Fall Fellows Retreat.
- Ganimian, A. J., A. D. Ho, and A. Campos Quintero (2026). The reliability of classroom observations and student surveys in non-research settings: Evidence from a middle-income country. (EdWorkingPaper No. 26-1442). Providence, RI: Annenberg Institute at Brown University. DOI: 10.26300/c9zp-p565.
- Ganimian, A. J., I. M. Mbiti, and A. Mishra (2026). Leveraging complementarities between teachers’ content knowledge and pedagogical supports: Experimental evidence from India. *Unpublished manuscript*. New York, NY: New York University (NYU).

- GEEAP (2020). Cost-effective approaches to improving global learning: What does recent evidence tell us are “Smart Buys” for improving learning in low- and middle-income countries? Washington, DC: World Bank; Foreign, Commonwealth & Development Office; Building Evidence in Education.
- Gilligan, D. O., N. Karachiwalla, I. Kasirye, A. M. Lucas, and D. Neal (2022). Educator incentives and educational triage in rural primary schools. *Journal of Human Resources* 57(1), 79–111.
- Glewwe, P., N. Ilias, and M. Kremer (2010). Teacher incentives. *American Economic Journal: Applied Economics* 2, 205–227.
- Glewwe, P., M. Kremer, and S. Moulin (2009). Many children left behind? Textbooks and test scores in Kenya. *American Economic Journal: Applied Economics* 1(1), 112–135.
- Hoge, R. D. and T. Coladarci (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research* 59(3), 297–313.
- IEA (2021). TIMSS 2023: Assessment frameworks. Edited by Mullis, I. V. S., Martin, M. O. & von Davier, M. TIMSS & PIRLS International Study Center. Lynch School of Education and Human Development, Boston College & International Association for the Evaluation of Educational Achievement (IEA).
- Kaufmann, E. (2024). Teachers’ judgment accuracy: A replication check by psychometric meta-analysis. *PLOS ONE* 19(7), e0307594.
- Kremer, M., C. Brannen, and R. Glennerster (2013). The challenge of education and learning in the developing world. *Science* 240, 297–300.
- Kremer, M., E. Miguel, and R. Thornton (2009). Incentives to learn. *The Review of Economics and Statistics* 91(3), 437–456.
- Ministry of Finance (2025). National budget speech 2025-26. Dhaka, Bangladesh: Government of the People’s Republic of Bangladesh. Retrieved from <https://nbr.gov.bd/uploads/budget/English.pdf>.
- Mizala, A. and M. Urquiola (2013). School markets: The impact of information approximating schools’ effectiveness. *Journal of Development Economics* 103, 313–335.
- Muralidharan, K., A. Singh, and A. J. Ganimian (2019). Disrupting education? Experimental evidence on technology-aided instruction in India. *American Economic Review* 109(4), 1–35.
- Muralidharan, K. and V. Sundararaman (2010). The impact of diagnostic feedback to teachers on student learning: Experimental evidence from india. *The Economic Journal* 120(F187-F203), F187–F203.
- Muralidharan, K. and V. Sundararaman (2011). Teacher performance pay: Experimental evidence from India. *Journal of Political Economy* 119(1), 39–77.

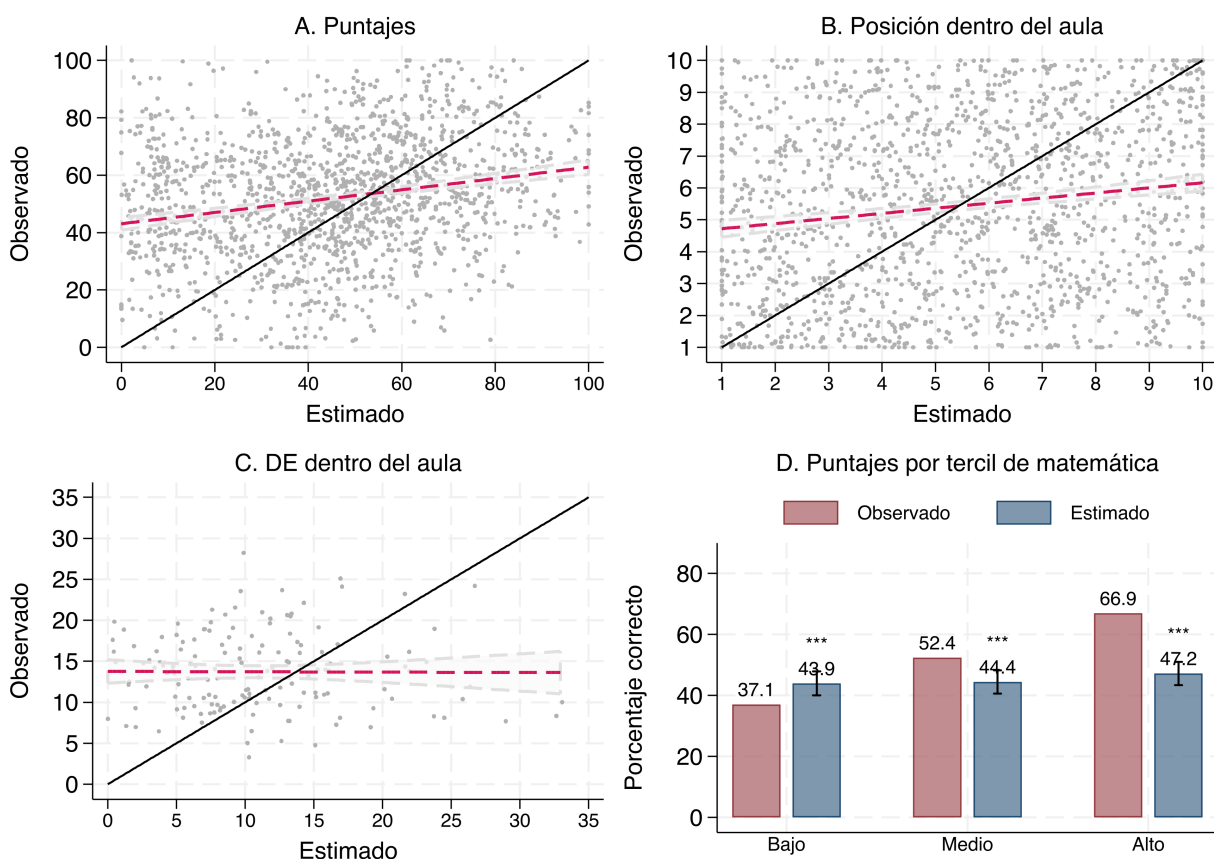
- Muralidharan, K. and Y. Zieleniak (2014). Chasing the syllabus: Measuring learning trajectories in developing countries with longitudinal data and item response theory. *Unpublished manuscript*. University of California, San Diego. San Diego, CA.
- NCTB (2013). Mathematics: Class six. Dhaka, Bangladesh: National Curriculum and Textbook Board (NCTB).
- NCTB (2021). National curriculum framework 2021. Dhaka, Bangladesh: National Curriculum and Textbook Board (NCTB).
- NCTB (2023). National curriculum framework 2021: Pre-primary to grade 12. Dhaka, Bangladesh: National Curriculum and Textbook Board (NCTB). Retrieved from <https://shed.portal.gov.bd/site/page/f33f6b5b-b650-4c9b-81ba-270ae88fa6e0/National-Curriculum-Framework-2021-%28English-Version%29>.
- Neilson, C., F. Gallego, and O. Molina (2015). The impact of information provision on human capital accumulation and child labor in peru. *Innovations for Poverty Action*. <https://www.poverty-action.org/printpdf/21321>.
- Nourani, V., N. Ashraf, and A. Banerjee (2025). Learning to teach by learning to learn. Working paper, February.
- Özler, B., L. C. H. Fernald, P. Kariger, C. McConnell, M. Neuman, and E. Fraga (2018). Combining pre-school teacher training with parenting education: A cluster-randomized controlled trial. *Journal of Development Economics* 133, 448–467.
- Piza, C., A. Zwager, M. Ruzzante, R. Dantas, and A. Loureiro (2024). Teacher-led innovations to improve education outcomes: Experimental evidence from Brazil. *Journal of Public Economics* 234, 105123.
- Premand, P. and O. Barry (2022). Behavioral change promotion, cash transfers and early childhood development: Experimental evidence from a government program in a low-income setting. *Journal of Development Economics* 158, 102921.
- Pritchett, L. and A. Beatty (2015). Slow down, you’re going too fast: Matching curricula to student skill levels. *International Journal of Educational Development* 40, 276–288.
- Rodriguez-Segura, D. (2024). NewGlobe policy reports. NewGlobe policy reports.
- Rodriguez-Segura, D. and I. Mbiti (2022). Back to the basics: Curriculum reform and student learning in tanzania.
- Sabarwal, S., M. Abu-Jawdeh, and R. Kapoor (2022). Teacher beliefs: Why they matter and what they are. *The World Bank Research Observer* 37(1), 73–106.
- Sankar, D. and T. Linden (2014). How much and what kind of teaching is there in elementary education in india? evidence from three states. Washington, DC: The World Bank.
- Shavelson, R. J. and N. M. Webb (1991). *Generalizability theory: A primer*. Sage.
- Singh, A. and P. Berg (2024). Myths of official measurement: Limits to test-based education reforms with weak governance. *Journal of Public Economics* 239, 105246.

- Stallings, J. A., S. L. Knight, and D. Markham (2014). Using the stallings observation system to investigate time on task in four countries. *Unpublished manuscript*. Washington, DC: The World Bank.
- Stallings, J. A. and G. G. Mohlman (1990). Issues in qualitative education research: Observation techniques. In H. J. Walberg & G. D. Haertel (Eds.) *The international encyclopedia of educational evaluation* (pp. 639-644). New York, NY: Pergamon Press.
- Südkamp, A., J. Kaiser, and J. Möller (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology* 104 (3), 743.
- Tatto, M. T., R. Peck, J. Schwille, K. Bankov, S. L. Senk, M. Rodriguez, L. Ingvarson, M. Reckase, and G. Rowley (2012). *Policy, practice, and readiness to teach primary and secondary mathematics in 17 countries: Findings from the IEA Teacher Education and Development Study in Mathematics (TEDS-M)*. Amsterdam, the Netherlands: Amsterdam, the Netherlands: International Evaluation Association (IEA).
- UIS (2026). UIS Data API. UNESCO Institute for Statistics (UIS) Data Browser. <https://api.uis.unesco.org/api/public/data/indicators>. Retrieved: May 21, 2026.
- UNESCO (2023). Efa global monitoring report 2023: Technology in education: A tool on whose terms?
- UNESCO (2026). Sustainable Development Goal 4 data. Paris, France: UNESCO Institute for Statistics. February 2026 data release.
- Wadmare, P., M. Nanda, R. Sabates, N. Sunder, and W. Wadhwa (2022). Understanding the accuracy of teachers' perceptions about low achieving learners in primary schools in rural india: An empirical analysis of alignments and misalignments. *International Journal of Educational Research Open* 3, 100198.
- Wolf, S., A. Ogan, K. K. Jasińska, S. Kembou, V. K. Cannanure, H. Whitehead, J. Corbett, F. Tanoh, B. Wortsman, J. Lim, H. Zhao, and A. Kumar (2026). Impacts of targeted instruction with mobile-based support on teachers and students in rural Côte d'Ivoire. *American Educational Research Journal* 63(4), 691–730.
- World Bank (2017). What is happening inside classrooms in Indian secondary schools? A time on task study in Madhya Pradesh and Tamil Nadu. Washington, DC: The World Bank and Educational Initiatives.
- World Bank (2018a). *World Development Report 2018: Learning to realize education's promise*. Washington, DC: The World Bank. Washington, DC: The World Bank.
- World Bank (2018b). *World Development Report 2018: Learning to realize education's promise*. Washington, DC: The World Bank.
- World Bank (2023). Learning Acceleration in Secondary Education Operation (P178487). Program Appraisal Document. <https://documents1.worldbank.org/curated/en/099090523122017246/pdf/B0SIB088cf52390990af10029ec20109467.pdf>.

World Bank (2025). Learning Acceleration in Secondary Education Operation (P178487). Implementation Status & Results Report, sequence no. 5, archived December 12, 2025. <https://documents1.worldbank.org/curated/en/099121225014523950/pdf/P178487-1b6e0dae-03a3-41aa-9ede-acf2f7aa620f.pdf>.

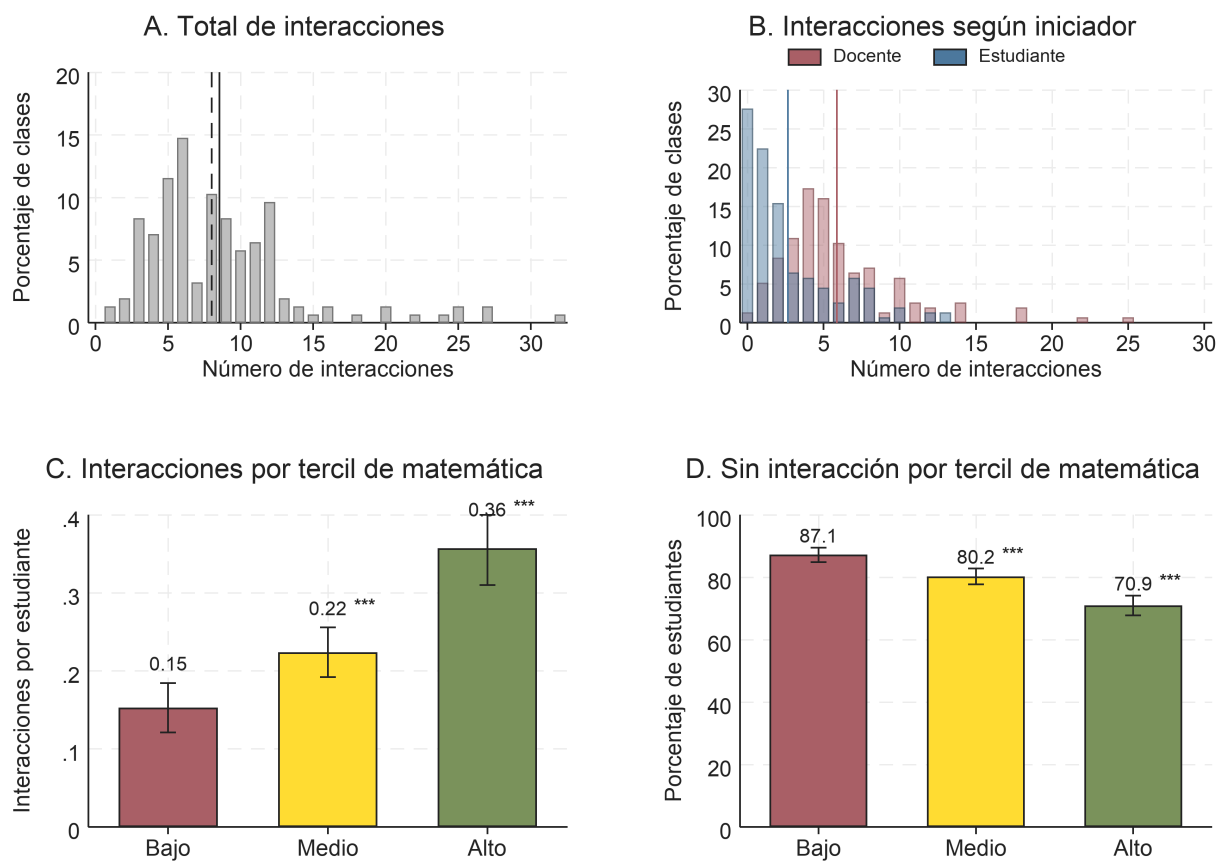
World Bank (2026). Population, total. <https://data.worldbank.org/indicator/SP.POP.TOTL>. Retrieved: May 21, 2026.

Figura 1: Estimaciones de los docentes sobre el rendimiento de los estudiantes (línea de base)



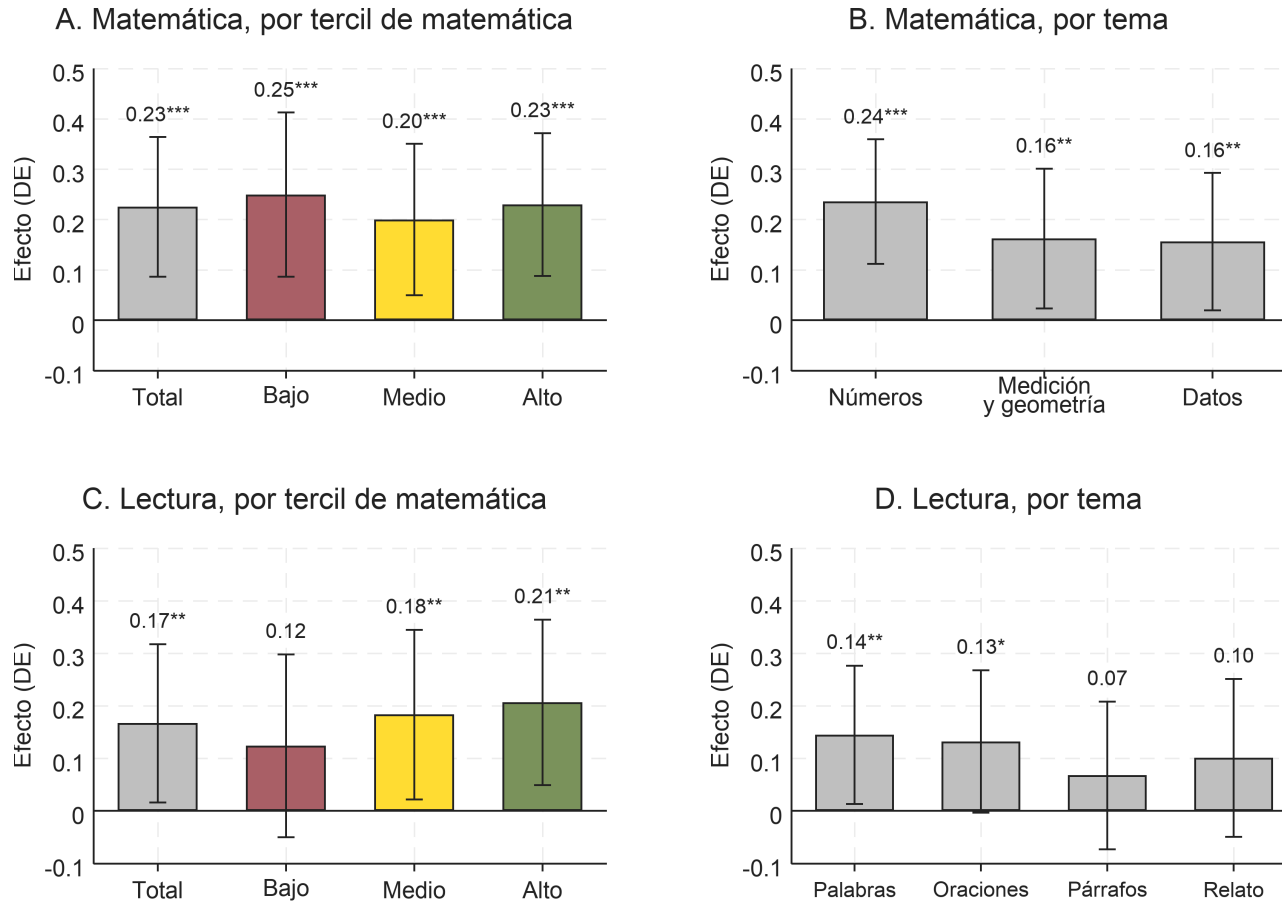
Notas: El panel A muestra la relación entre los puntajes reales de los estudiantes en las evaluaciones de la intervención y los puntajes estimados por los docentes en la prueba de matemática de línea de base. Los paneles B y C muestran las relaciones entre los ordenamientos reales y estimados de los estudiantes dentro del aula y entre los DE reales y estimados dentro del aula. El panel D compara los puntajes reales de las evaluaciones de la intervención y los porcentajes de respuestas correctas estimados según el tercil de rendimiento en matemática de línea de base dentro del aula de los estudiantes en las evaluaciones de la intervención. Los paneles A, B y D están a nivel de diada estudiante-docente; el panel C está a nivel de docente. La línea roja punteada es un ajuste lineal y la línea negra es la línea de 45 grados. En el panel D, las barras son medias ajustadas por estrato, y los bigotes alrededor de las barras estimadas muestran intervalos de confianza del 95 % para las diferencias entre lo estimado y lo real de especificaciones con efectos fijos de estrato de aleatorización y errores estándar agrupados a nivel de escuela. Las correlaciones entre los valores reales y los estimados son de 0.23 en el panel A, 0.16 en el panel B y -0.01 en el panel C. La figura incluye a los estudiantes del grupo de tratamiento con puntajes en las evaluaciones de la intervención en la línea de base.

Figura 2: Interacciones entre estudiantes y docentes en el grupo de control (línea de cierre)



Notas: Los paneles A y B muestran la distribución de las interacciones entre estudiantes y docentes por clase en las escuelas de control. Las líneas continuas marcan las medias, y la línea punteada del panel A marca la mediana. Los paneles C y D reportan medias ajustadas de los estudiantes de control según el tercil de rendimiento en matemática de línea de base dentro del aula en las evaluaciones de impacto, provenientes de regresiones agrupadas del grupo de control con variables indicadoras de tercil y efectos fijos de estrato de aleatorización. Los bigotes muestran intervalos de confianza del 95%; los asteriscos indican pruebas contra el grupo de bajo rendimiento.

Figura 3: Impacto sobre el rendimiento de los estudiantes (línea de cierre)



Notas: Las barras muestran los efectos de tratamiento; los bigotes muestran intervalos de confianza del 95%. Los paneles A y C reportan los efectos globales y por tercil de rendimiento en matemática de línea de base dentro del aula en las evaluaciones de impacto. Los paneles B y D reportan los efectos promedio por tema de matemática y por tema de lectura, respectivamente. Todos los resultados son puntajes escalados por TRI estandarizados para tener una media de cero y un DE de uno en el grupo de control global. Todas las especificaciones incluyen efectos fijos de estrato de aleatorización y covariables de línea de base seleccionadas mediante LASSO. Los errores estándar están agrupados a nivel de escuela. * significativo al 10%; ** significativo al 5%; *** significativo al 1%.

Tabla 1: Cambio en las evaluaciones de la intervención según contenidos y estudiantes señalados entre las escuelas de tratamiento (de línea de base a línea intermedia)

	(1) Media de línea de base	(2) Media de línea intermedia	(3) Cambio	(4) N (observaciones de contenido-clase/estudiante)
<i>A. Temas señalados vs. no señalados</i>				
Temas				
(1) Temas no señalados	55.5	65.3	9.80	312
(2) Tema señalado	21.5	46.5	25.0	156
Fila (2) – (1)			14.5*** (2.71)	468
Subtemas				
(1) Subtemas no señalados	54.6	65.3	10.8	1,248
(2) Subtemas señalados	4.52	42.2	37.6	156
Fila (2) – (1)			19.4*** (3.38)	1,404
Ítems según umbrales de color				
(1) Ítems verdes	86.0	83.0	-2.92	681
(2) Ítems rojos/amarillos	25.4	51.2	25.8	879
Fila (2) – (1)			34.7*** (2.48)	1,560
(3) Ítems amarillos	50.6	64.8	14.2	310
(4) Ítems rojos	11.7	43.8	32.2	569
Fila (4) – (3)			28.1*** (2.90)	879
<i>B. Estudiantes señalados vs. no señalados</i>				
(1) Estudiantes no señalados	54.3	67.0	12.7	4,154
(2) Estudiantes señalados	40.7	63.3	22.5	809
Fila (2) – (1)			9.85*** (0.94)	4,963
<i>C. Estudiantes según la cantidad de temas señalados</i>				
(0) 0 temas	54.3	67.0	12.7	4,154
(1) 1 tema	42.6	63.8	21.2	606
(2) 2 temas	36.4	61.9	25.5	176
(3) 3 temas	26.3	59.6	33.3	27
Fila (1) – (0)			8.51*** (1.04)	4,963
Fila (2) – (0)			12.8*** (1.72)	4,963
Fila (3) – (0)			20.6*** (5.08)	4,963
<i>D. Temas señalados vs. no señalados entre los estudiantes señalados</i>				
(1) Temas no señalados	49.4	59.2	9.79	1,388
(2) Temas señalados	13.1	54.9	41.8	958
Fila (2) – (1)			32.9*** (1.70)	2,346

Notas: La tabla utiliza las clases y los estudiantes de tratamiento con puntajes en las evaluaciones de la intervención en la línea de base y en la línea intermedia. El panel A compara los temas señalados y no señalados; el panel B compara los estudiantes señalados y no señalados; el panel C compara a los estudiantes según la cantidad de temas en los que fueron señalados; y el panel D compara los temas señalados y no señalados entre los estudiantes señalados. Los colores de los ítems se basan en las medias de línea de base: rojo <33 %, amarillo del 33 % a <66 %, y verde ≥66 %. Las comparaciones de contenidos incluyen efectos fijos de clase y de tema/subtema/ítem; el panel D usa efectos fijos de estudiante y de tema. Los puntajes, los cambios y las diferencias están en p.p.; los errores estándar agrupados a nivel de escuela aparecen entre paréntesis. * 10 %; ** 5 %; *** 1 %.

Tabla 2: Impacto sobre el rendimiento en matemática (línea de cierre)

	(1) Todos los estudiantes	(2) Bajo rendimiento	(3) Rendimiento medio	(4) Alto rendimiento	(5) p (2)=(3)	(6) p (2)=(4)
<i>A. Global</i>						
<u>Global</u>	0.23*** (0.07)	0.25*** (0.08)	0.20*** (0.08)	0.23*** (0.07)	0.29	0.54
Media de control	0.00	-0.30	-0.01	0.32		
<i>B. Por tema</i>						
<u>Número</u>	0.24*** (0.06)	0.27*** (0.08)	0.22*** (0.07)	0.23*** (0.07)	0.41	0.55
Media de control	0.00	-0.30	-0.01	0.32		
<u>Medición y geometría</u>	0.16** (0.07)	0.20** (0.08)	0.12 (0.08)	0.17** (0.07)	0.12	0.40
Media de control	0.00	-0.24	0.00	0.25		
<u>Datos</u>	0.16** (0.07)	0.13 (0.08)	0.16** (0.08)	0.18** (0.08)	0.67	0.55
Media de control	0.00	-0.17	-0.03	0.21		
p (igualdad entre temas)	0.13					
<i>C. Por habilidad</i>						
<u>Conocer</u>	0.18** (0.07)	0.21** (0.08)	0.17** (0.08)	0.15** (0.07)	0.43	0.30
Media de control	0.00	-0.27	-0.01	0.30		
<u>Aplicar</u>	0.19*** (0.06)	0.21*** (0.08)	0.17** (0.07)	0.22*** (0.06)	0.59	0.86
Media de control	0.00	-0.26	-0.01	0.27		
<u>Razonar</u>	0.26*** (0.07)	0.28*** (0.08)	0.24*** (0.08)	0.27*** (0.07)	0.36	0.63
Media de control	0.00	-0.26	-0.01	0.28		
p (igualdad entre habilidades)	0.14					
N (estudiantes)	4,542	1,524	1,576	1,442		

Notas: La tabla muestra el impacto de la intervención sobre la prueba de matemática de cierre, en forma global (panel A), por tema (panel B) y por habilidad cognitiva (panel C). Los puntajes escalados por TRI están estandarizados para tener una media de cero y un DE de uno en el grupo de control global; las medias de control son por lo tanto cero para la columna 1 por construcción y negativas (positivas) para los estudiantes de bajo (alto) rendimiento. La columna 1 muestra los resultados de todos los estudiantes, y las columnas 2-4 los de los estudiantes de los terciles inferior, medio y superior de rendimiento en matemática de línea de base dentro del aula en las evaluaciones de impacto, respectivamente. Las columnas 5 y 6 reportan los valores p de las pruebas de igualdad de los efectos de tratamiento de las columnas 2 y 3, y de las columnas 2 y 4, respectivamente. Las filas p (igualdad entre temas) y p (igualdad entre habilidades) reportan pruebas ómnibus de que los tres efectos de tratamiento del panel correspondiente son iguales, estimadas con regresiones apiladas con efectos fijos de estrato de aleatorización específicos por dominio y controles de línea de base. Todas las especificaciones usan errores estándar agrupados a nivel de escuela; las especificaciones de efectos de tratamiento incluyen covariables de línea de base seleccionadas mediante LASSO. Los errores estándar aparecen entre paréntesis. * significativo al 10 %; ** significativo al 5 %; *** significativo al 1 %.

Tabla 3: Impacto sobre el rendimiento en lectura (línea de cierre)

	(1) Todos los estudiantes	(2) Bajo rendimiento	(3) Rendimiento medio	(4) Alto rendimiento	(5) p (2)=(3)	(6) p (2)=(4)
<i>A. Global</i>						
<u>Lectura</u>	0.17** (0.08)	0.12 (0.09)	0.18** (0.08)	0.21** (0.08)	0.46	0.39
Media de control	0.00	-0.19	-0.05	0.24		
<i>B. Por nivel de insumo</i>						
<u>Palabras</u>						
Tratamiento	0.15** (0.07)	0.15* (0.08)	0.15* (0.08)	0.13** (0.07)	0.94	0.76
Media de control	0.00	-0.21	-0.02	0.24		
<u>Oraciones</u>						
Tratamiento	0.13* (0.07)	0.09 (0.08)	0.14* (0.08)	0.18** (0.08)	0.57	0.40
Media de control	0.00	-0.13	-0.04	0.17		
<u>Párrafos</u>						
Tratamiento	0.07 (0.07)	-0.01 (0.08)	0.11 (0.08)	0.12 (0.08)	0.12	0.13
Media de control	0.00	-0.12	-0.06	0.19		
<u>Relato</u>						
Tratamiento	0.10 (0.08)	0.07 (0.09)	0.09 (0.08)	0.16* (0.09)	0.91	0.28
Media de control	0.00	0.00	-0.02	0.02		
<i>C. Por habilidad</i>						
<u>Vocabulario</u>						
Tratamiento	0.15** (0.07)	0.13 (0.08)	0.15* (0.08)	0.16** (0.07)	0.89	0.90
Media de control	0.00	-0.22	-0.03	0.26		
<u>Comprensión</u>						
Tratamiento	0.14* (0.08)	0.07 (0.09)	0.17** (0.08)	0.20** (0.09)	0.21	0.15
Media de control	0.00	-0.11	-0.05	0.17		
N (estudiantes)	4,542	1,524	1,576	1,442		

Notas: La tabla muestra el impacto de la intervención sobre la prueba de lectura de cierre, en forma global (panel A), por nivel de insumo (panel B) y por habilidad lectora (panel C). Los insumos (Palabras, Oraciones, Párrafos, Relato) y las habilidades (Vocabulario, Comprensión) siguen el marco de evaluación de lectura, en el que cada ítem se clasifica según ambos. Los puntajes escalados por TRI están estandarizados para tener una media de cero y un DE de uno en el grupo de control global; las medias de control son por lo tanto cero para la columna 1 por construcción y negativas (positivas) para los estudiantes de bajo (alto) rendimiento. La columna 1 muestra los resultados de todos los estudiantes, y las columnas 2-4 los de los estudiantes de los terciles inferior, medio y superior de rendimiento en matemática de línea de base dentro del aula en las evaluaciones de impacto, respectivamente. Las columnas 5 y 6 reportan los valores p de las pruebas de igualdad de los efectos de tratamiento de las columnas 2 y 3, y de las columnas 2 y 4, respectivamente. Todas las especificaciones incluyen efectos fijos de estrato de aleatorización y covariables de línea de base seleccionadas mediante LASSO. Los errores estándar agrupados a nivel de escuela aparecen entre paréntesis. * significativo al 10%; ** significativo al 5%; *** significativo al 1%.

Tabla 4: Impacto sobre las habilidades socioemocionales (línea de cierre)

	(1) Todos los estudiantes	(2) Bajo rendimiento	(3) Rendimiento medio	(4) Alto rendimiento	(5) p (2)=(3)	(6) p (2)=(4)
Autoeficacia	0.14*** (0.05)	0.14** (0.07)	0.18*** (0.06)	0.09 (0.07)	0.51	0.53
Autogestión	0.05 (0.05)	-0.01 (0.07)	0.09 (0.06)	0.06 (0.07)	0.15	0.36
Mentalidad de crecimiento	0.11** (0.05)	0.20*** (0.07)	0.09 (0.06)	0.04 (0.06)	0.16	0.03
Miedo al fracaso	-0.03 (0.05)	0.02 (0.07)	-0.07 (0.06)	-0.04 (0.07)	0.19	0.44
Orientación a metas de tarea	0.10** (0.05)	0.13* (0.07)	0.10 (0.06)	0.07 (0.06)	0.65	0.42
Motivación por dominar tareas	0.03 (0.05)	0.03 (0.06)	0.06 (0.06)	-0.02 (0.06)	0.69	0.46
Determinación	0.06 (0.05)	0.07 (0.07)	0.09 (0.06)	0.01 (0.07)	0.83	0.35
N (estudiantes)	4,542	1,524	1,576	1,442		

Notas: La tabla muestra el impacto de la intervención sobre siete habilidades socioemocionales autorreportadas medidas en la línea de cierre. Los modelos de TRI de respuesta graduada agrupan los ítems de la línea de base y de la de cierre; los puntajes están estandarizados para tener una media de cero y un DE de uno en el grupo de control en cada ronda. La columna 1 muestra los resultados de todos los estudiantes, y las columnas 2-4 los de los estudiantes de los terciles inferior, medio y superior de rendimiento en matemática de línea de base dentro del aula en las evaluaciones de impacto, respectivamente. Las columnas 5 y 6 reportan los valores p de las pruebas de igualdad de los efectos de tratamiento de las columnas 2 y 3, y de las columnas 2 y 4, respectivamente. Todas las especificaciones incluyen efectos fijos de estrato de aleatorización y covariables de línea de base seleccionadas mediante LASSO. Los errores estándar agrupados a nivel de escuela van entre paréntesis. * 10%; ** 5%; *** 1%.

Tabla 5: Impacto en las estimaciones docentes de los puntajes estudiantiles (línea de cierre)

	(1) Todos los estudiantes	(2) Bajo rendimiento	(3) Rendimiento medio	(4) Alto rendimiento	(5) p (2)=(3)	(6) p (2)=(4)
<i>A. Precisión</i>						
<u>Brecha absoluta de puntaje (DE)</u>	0.14 (0.20)	0.11 (0.19)	0.20 (0.24)	0.09 (0.19)	0.50	0.90
Media de control	1.74	1.83	1.77	1.62		
<u>Brecha absoluta de ordenamiento (percentiles)</u>	-0.01 (0.01)	0.01 (0.01)	-0.01 (0.01)	-0.02 (0.01)	0.42	0.08
Media de control	0.31	0.31	0.31	0.29		
<i>B. Dirección</i>						
<u>Brecha de puntaje (DE)</u>	-0.47* (0.24)	-0.48* (0.25)	-0.48* (0.29)	-0.42* (0.23)	1.00	0.69
Media de control	-0.26	-0.11	-0.33	-0.35		
<u>Brecha de ordenamiento (percentiles)</u>	-0.01* (0.01)	-0.01 (0.02)	-0.01 (0.02)	-0.01 (0.02)	0.96	0.95
Media de control	0.00	0.04	0.00	-0.04		
<u>El docente sobreestima</u>	-0.04 (0.04)	-0.03 (0.04)	-0.04 (0.04)	-0.05 (0.04)	0.88	0.61
Media de control	0.46	0.50	0.46	0.40		
<u>El docente subestima</u>	0.03 (0.04)	0.03 (0.04)	0.04 (0.04)	0.03 (0.04)	0.72	1.00
Media de control	0.50	0.46	0.50	0.56		
<i>C. Dispersión</i>						
<u>Brecha de DE (puntos)</u>	0.07 (0.72)	0.00 (0.73)	-0.01 (0.74)	0.17 (0.73)	0.97	0.50
Media de control	-0.63	-0.50	-0.71	-0.69		
N (días estudiante-docente)	4,314	1,423	1,503	1,388		

Notas: Impacto sobre las estimaciones de los docentes del desempeño de los estudiantes en matemática en la línea de cierre. El panel A muestra las brechas absolutas de puntajes (normalizadas por el DE dentro del aula de los puntajes reales) y las brechas absolutas de ordenamiento (rango percentilar dentro del aula). El panel B muestra las brechas de puntaje y de ordenamiento con signo y variables indicadoras de sobreestimación y subestimación. El panel C muestra la brecha de DE. Las brechas con signo negativo (positivo) indican subestimación (sobreestimación). Las cols. 2-4 son terciles dentro del aula de rendimiento en matemática de línea de base en las evaluaciones de impacto; las cols. 5-6 dan los valores p de las cols. 2 vs. 3 y 2 vs. 4. Todas las especificaciones incluyen EF de estrato y covariables de línea de base seleccionadas por LASSO. Los EE agrupados a nivel de escuela entre paréntesis. * 10%; ** 5%; *** 1%.

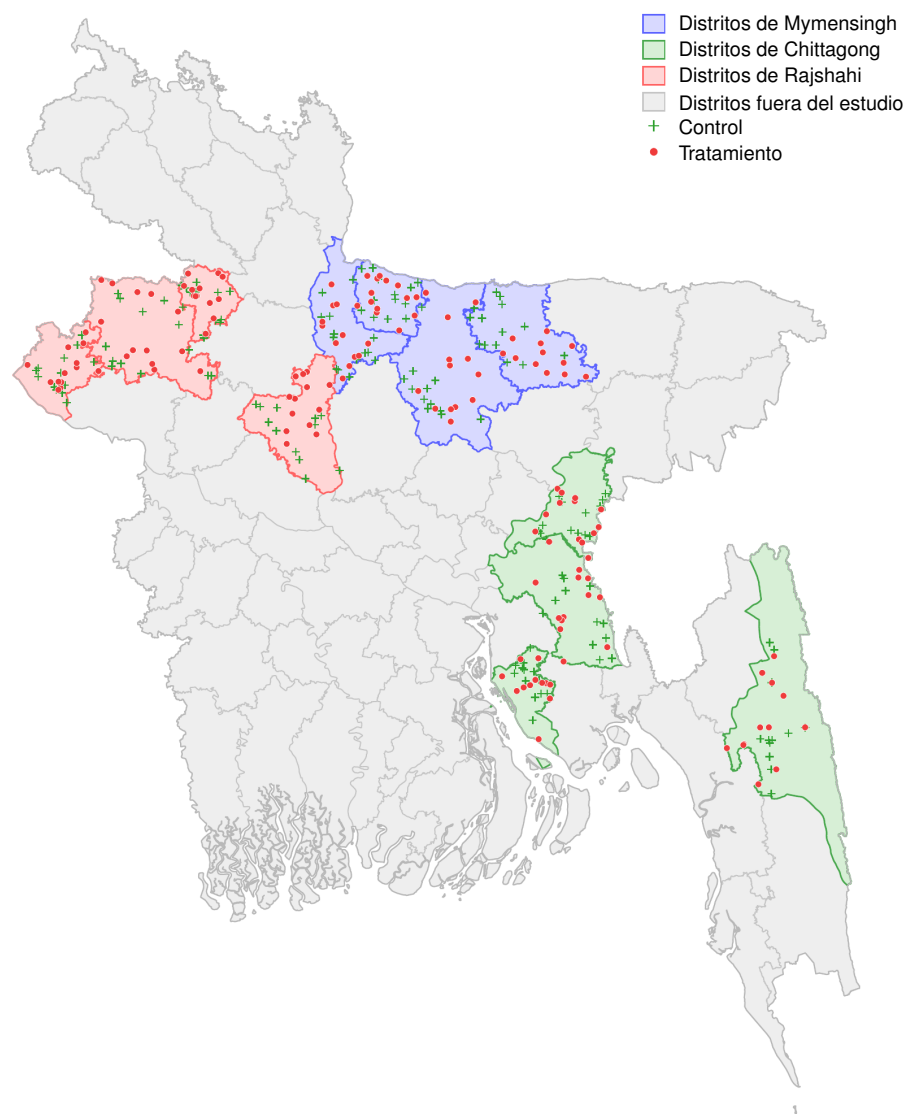
Tabla 6: Impacto sobre las interacciones entre estudiantes y docentes (línea de cierre)

	(1) Todos los estudiantes	(2) Bajo rendimiento	(3) Rendimiento medio	(4) Alto rendimiento	(5) p (2)=(3)	(6) p (2)=(4)
<i>A. Global</i>						
<u>Cantidad de interacciones</u>	-0.02 (0.02)	0.01 (0.02)	-0.02 (0.02)	-0.07* (0.04)	0.22	0.04
Media de control	0.24	0.15	0.22	0.36		
<u>Proporción con 0 interacciones</u>	0.02 (0.01)	-0.02 (0.02)	0.02 (0.02)	0.05** (0.02)	0.10	0.02
Media de control	0.80	0.87	0.80	0.71		
<i>B. Según quién inició</i>						
<u>Docente</u>	-0.02 (0.02)	0.00 (0.02)	-0.01 (0.02)	-0.04 (0.03)	0.67	0.39
Media de control	0.16	0.11	0.15	0.24		
<u>Estudiante</u>	0.00 (0.01)	0.02 (0.01)	-0.01 (0.01)	-0.03 (0.02)	0.15	0.02
Media de control	0.08	0.04	0.07	0.12		
<i>C. Por tipo</i>						
<u>Afirmación</u>	-0.01 (0.01)	0.00 (0.02)	0.00 (0.01)	-0.04 (0.02)	0.95	0.14
Media de control	0.08	0.06	0.06	0.13		
<u>Pregunta</u>	-0.01 (0.02)	0.02 (0.02)	-0.01 (0.02)	-0.03 (0.03)	0.11	0.13
Media de control	0.16	0.10	0.16	0.23		
<i>D. Por duración</i>						
<u>< 5 segundos</u>	-0.01 (0.01)	0.01 (0.01)	0.00 (0.01)	-0.03* (0.02)	0.27	< 0.01
Media de control	0.04	0.02	0.04	0.06		
<u>6-10 segundos</u>	0.00 (0.01)	0.01 (0.01)	0.00 (0.01)	-0.01 (0.02)	0.50	0.57
Media de control	0.08	0.05	0.08	0.12		
<u>> 10 segundos</u>	-0.02 (0.01)	-0.01 (0.01)	-0.01 (0.02)	-0.03 (0.03)	0.62	0.41
Media de control	0.12	0.08	0.11	0.17		
N (estudiantes)	5,920	2,077	2,058	1,785		

Notas: La tabla muestra el impacto de la intervención sobre la cantidad de interacciones entre estudiantes y docentes por estudiante por clase observadas en la línea de cierre, en forma global (panel A), según quién inició la interacción (panel B), por tipo de interacción (panel C) y por duración de la interacción (panel D). Los resultados son conteos observados por estudiante durante una única clase de observación de aula. La columna 1 muestra los resultados de todos los estudiantes, y las columnas 2-4 los de los estudiantes de los terciles inferior, medio y superior de rendimiento en matemática de línea de base dentro del aula en las evaluaciones de impacto, respectivamente. Las columnas 5 y 6 reportan los valores p de las pruebas de igualdad de los efectos de tratamiento de las columnas 2 y 3, y de las columnas 2 y 4, respectivamente. Todas las especificaciones incluyen efectos fijos de estrato de aleatorización y covariables de línea de base seleccionadas mediante LASSO. Los errores estándar agrupados a nivel de escuela aparecen entre paréntesis. * significativo al 10%; ** significativo al 5%; *** significativo al 1%.

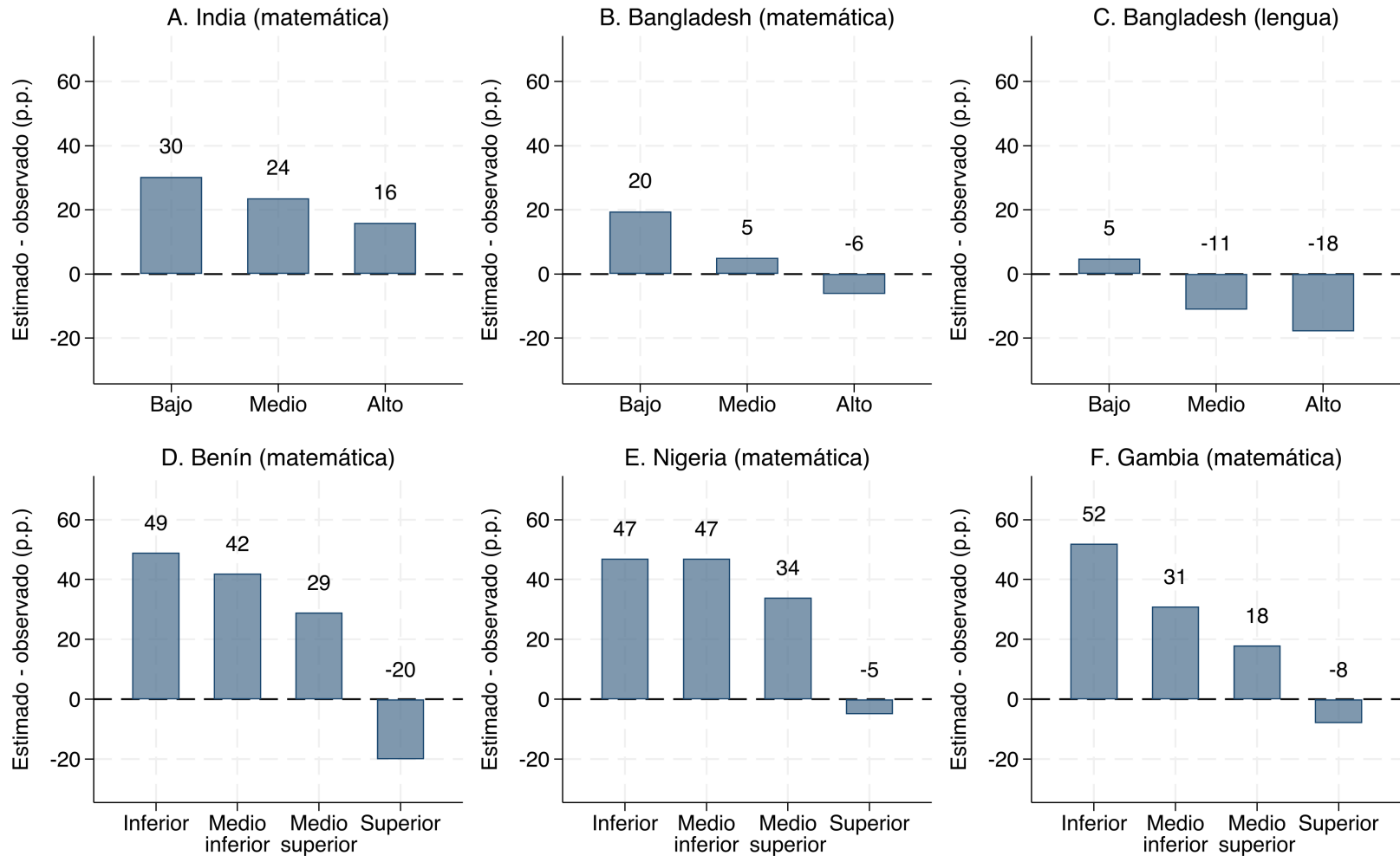
Apéndice A Figuras y tablas adicionales

Figura A.1: Distribución espacial de las escuelas del estudio



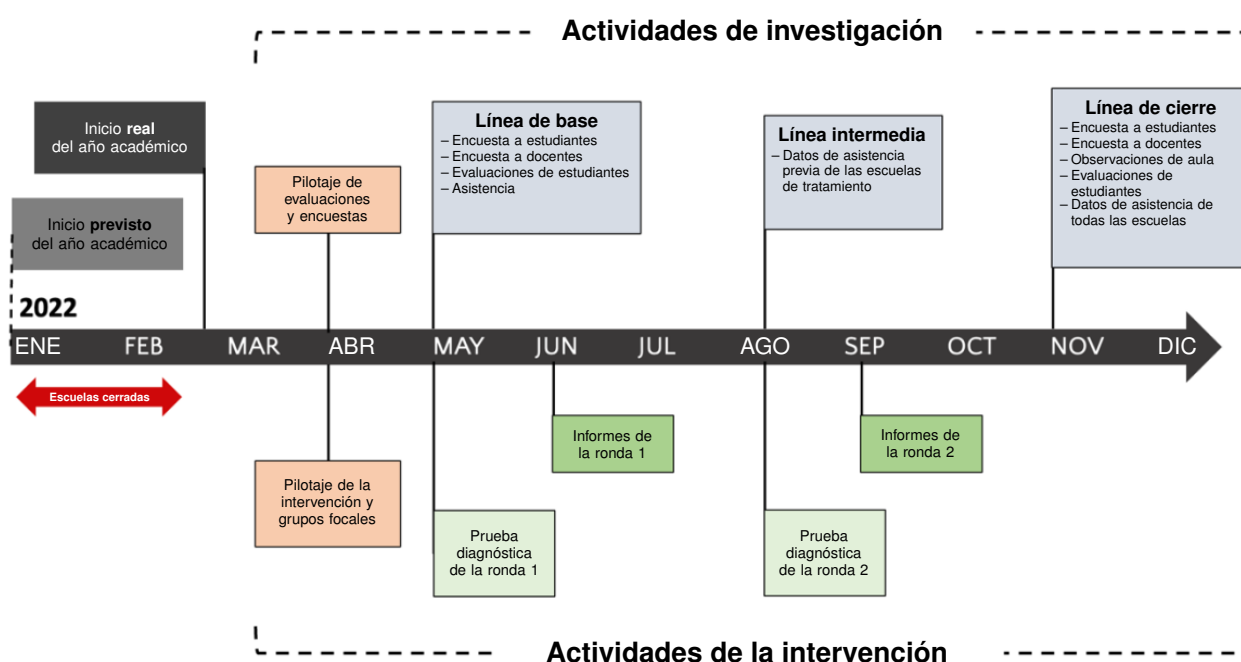
Notas: La figura muestra la ubicación georreferenciada de todas las escuelas de la evaluación de impacto en Chittogram (antes Chittagong) en verde, Mymensingh en azul y Rajshahi en rojo.

Figura A.2: Evidencia externa sobre las estimaciones de los docentes del rendimiento de los estudiantes (línea de base)



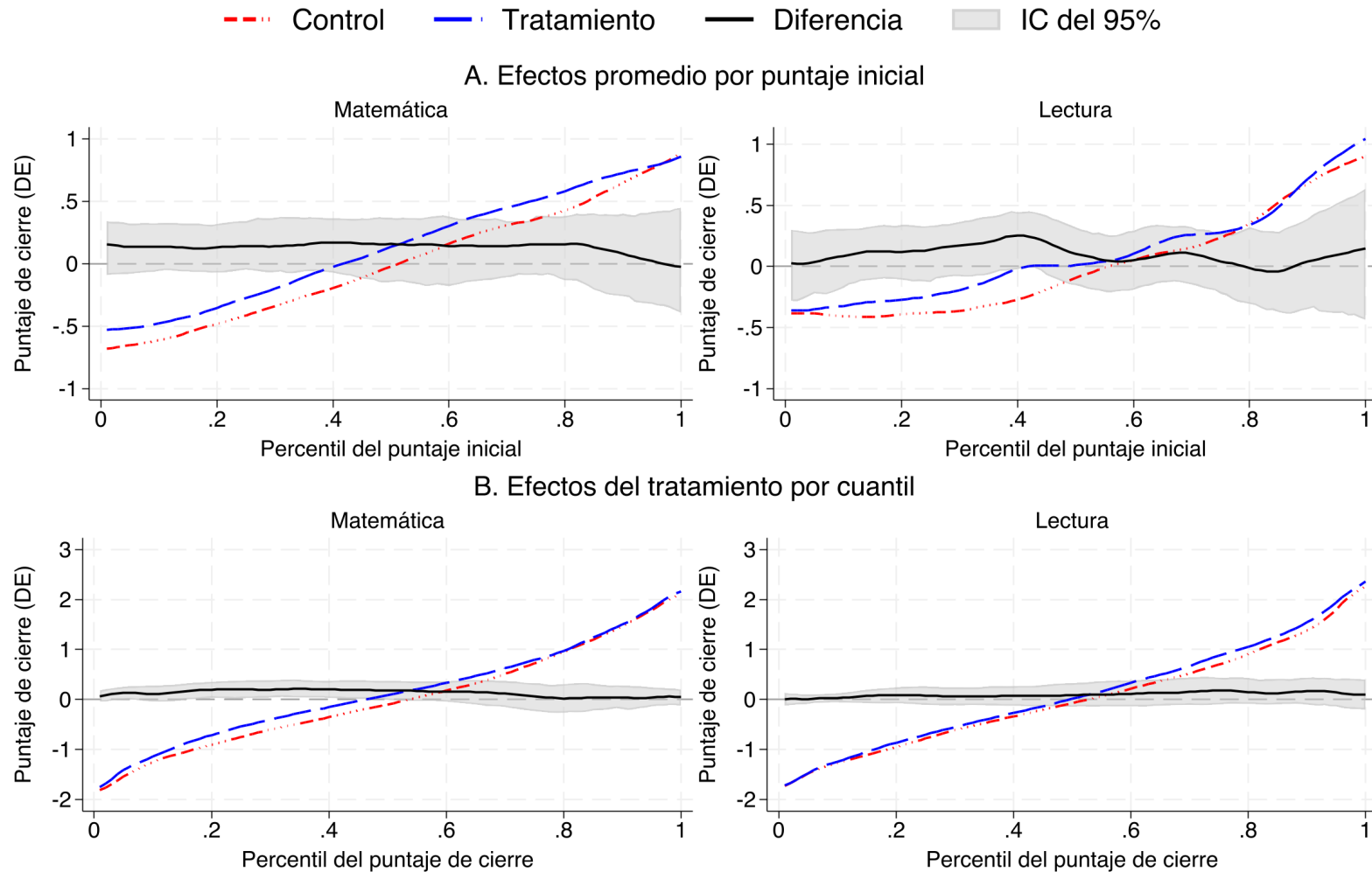
Notas: Los paneles A-C muestran las brechas entre lo estimado y lo real en los porcentajes de respuestas correctas de los estudiantes según su tercil de rendimiento dentro del aula en matemática en India, matemática en Bangladesh y lengua en Bangladesh, usando los valores de la Tabla A.6 de Djaker et al. (2024). Los paneles D-F muestran las brechas entre lo estimado y lo real en las creencias de los directores sobre el porcentaje de estudiantes que podría responder correctamente un problema de suma, por cuartil de desempeño escolar, en Benín, Nigeria y Gambia, usando los gráficos de Rodríguez-Segura (2024). Los valores positivos indican sobreestimación.

Figura A.3: Cronograma de las actividades de investigación y de la intervención



Notas: La figura muestra el cronograma de las actividades de investigación y de la intervención durante el año escolar 2022. Las escuelas reabrieron el 22 de febrero de 2022, después de los cierres debidos al COVID-19. Marzo y abril se utilizaron para pilotear la intervención, las evaluaciones y las encuestas.

Figura A.4: Efectos distribucionales del tratamiento sobre el rendimiento de los estudiantes (línea de cierre)



Notas: La figura muestra los efectos distribucionales del tratamiento sobre el rendimiento en matemática (columna izquierda) y en lectura (columna derecha) en la línea de cierre. En cada panel, las líneas roja y azul son regresiones de polinomios locales de los puntajes estandarizados de cierre sobre el rango percentilar dentro del grupo de los puntajes de línea de base (fila superior) o de los puntajes de cierre (fila inferior), estimadas por separado para los estudiantes de control y de tratamiento; la línea negra es su diferencia, y la banda sombreada es su intervalo de confianza del 95 %, construido mediante un bootstrap por conglomerados de 200 iteraciones (conglomerados de escuela, estratificados por tratamiento). Los puntajes están estandarizados para tener una media de cero y un DE de uno en el grupo de control.

Tabla A.1: Comparación de escuelas dentro y fuera de la muestra: evaluación nacional de estudiantes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Todas las divisiones				Divisiones en la muestra			
	Todas las escuelas	Escuelas fuera de la muestra	Escuelas en la muestra	Col.(3)-(2)	Todas las escuelas	Escuelas fuera de la muestra	Escuelas en la muestra	Col.(7)-(6)
<i>A. Matemática</i>								
Total (est.)	0.00	0.00	-0.19	-0.20	0.04	0.05	-0.19	-0.25
	[0.95]	[0.95]	[0.81]	(0.15)	[0.96]	[0.96]	[0.81]	(0.15)
Números (pct.)	0.61	0.61	0.58	-0.04	0.62	0.63	0.58	-0.05
	[0.23]	[0.23]	[0.20]	(0.03)	[0.23]	[0.23]	[0.20]	(0.03)
Medición (pct.)	0.55	0.56	0.50	-0.06	0.57	0.57	0.50	-0.07
	[0.27]	[0.27]	[0.26]	(0.04)	[0.28]	[0.28]	[0.26]	(0.04)
Datos (pct.)	0.64	0.64	0.61	-0.02	0.64	0.64	0.61	-0.03
	[0.26]	[0.26]	[0.27]	(0.04)	[0.26]	[0.26]	[0.27]	(0.04)
Geometría (pct.)	0.52	0.53	0.45	-0.07*	0.53	0.54	0.45	-0.08**
	[0.25]	[0.25]	[0.23]	(0.04)	[0.25]	[0.25]	[0.23]	(0.04)
Álgebra (pct.)	0.48	0.48	0.48	-0.01	0.49	0.49	0.48	-0.02
	[0.28]	[0.28]	[0.25]	(0.04)	[0.28]	[0.28]	[0.25]	(0.04)
N (estudiantes)	28,566	27,802	470	28,566	10,926	10,456	470	10,926
<i>B. Bangla</i>								
Total (est.)	0.00	0.00	-0.14	-0.14	0.03	0.04	-0.14	-0.18
	[0.94]	[0.94]	[0.90]	(0.12)	[0.91]	[0.91]	[0.90]	(0.13)
Lectura (pct.)	0.63	0.63	0.60	-0.03	0.64	0.64	0.60	-0.03
	[0.19]	[0.19]	[0.19]	(0.02)	[0.18]	[0.18]	[0.19]	(0.03)
Vocabulario (pct.)	0.81	0.81	0.80	-0.01	0.81	0.81	0.80	-0.02
	[0.20]	[0.20]	[0.20]	(0.02)	[0.20]	[0.20]	[0.20]	(0.02)
Gramática (pct.)	0.78	0.78	0.75	-0.04	0.79	0.79	0.75	-0.04*
	[0.22]	[0.22]	[0.22]	(0.02)	[0.22]	[0.22]	[0.22]	(0.03)
N (estudiantes)	28,566	27,802	470	28,566	10,926	10,456	470	10,926

Notas: La tabla muestra las medias y los DE de todas las escuelas fuera de la muestra y de las escuelas dentro de la muestra, y la diferencia entre estos grupos en la evaluación nacional de estudiantes de Bangladesh, la Evaluación Nacional de Estudiantes de Secundaria (NASS). Las columnas 1-4 reportan comparaciones nacionales, mientras que las columnas 5-8 restringen la muestra a las tres divisiones incluidas en el experimento. El panel A muestra los resultados de matemática y el panel B los de bangla (lengua). * significativo al 10 %; ** significativo al 5 %; *** significativo al 1 %.

Tabla A.2: Pruebas de balance de características y habilidades estudiantiles (línea de base)

Variable	(1) Control	(2) Tratamiento	(3) Col.(2)-(1)
<i>A. Características de contexto</i>			
Mujer	0.61 [0.49]	0.65 [0.48]	0.04* (0.03)
Edad	12.0 [0.99]	12.0 [0.95]	-0.04 (0.04)
Habla bangla en el hogar	0.94 [0.24]	0.94 [0.23]	0.00 (0.01)
Madre completó la secundaria	0.45 [0.50]	0.40 [0.49]	-0.05*** (0.02)
Activos del hogar (est.)	0.00 [1.00]	0.13 [0.99]	0.14*** (0.05)
N (estudiantes)	2,966	2,974	5,940
<i>B. Habilidades socioemocionales</i>			
Mentalidad de crecimiento (est.)	0.00 [1.00]	0.06 [1.00]	0.05 (0.04)
Autogestión (est.)	0.00 [1.00]	0.02 [0.99]	0.02 (0.04)
Autoeficacia (est.)	0.00 [1.00]	0.07 [0.98]	0.07 (0.04)
Orientación a metas de tarea (est.)	0.00 [1.00]	0.00 [0.95]	0.00 (0.03)
Motivación por dominar tareas (est.)	0.00 [1.00]	-0.01 [0.99]	-0.01 (0.04)
Miedo al fracaso (est.)	0.00 [1.00]	-0.03 [1.03]	-0.03 (0.04)
Determinación (est.)	0.00 [1.00]	0.03 [1.02]	0.03 (0.04)
N (estudiantes)	2,966	2,974	5,940
<i>C. Habilidades académicas</i>			
Matemática (TRI, est.)	0.00 [1.00]	-0.17 [1.01]	-0.18** (0.07)
Lectura (TRI, est.)	0.00 [1.00]	-0.13 [1.00]	-0.14* (0.08)
Inteligencia fluida (est.)	0.00 [1.00]	-0.06 [1.03]	-0.06 (0.05)
N (estudiantes)	2,966	2,974	5,940

Notas: La tabla muestra las medias y los DE de los estudiantes de los grupos de control (columna 1) y de tratamiento (columna 2). También prueba las diferencias entre grupos, teniendo en cuenta los efectos fijos de estrato de aleatorización (columna 3). El panel A muestra los resultados de las características de contexto, el panel B los de las habilidades socioemocionales y el panel C los de las habilidades académicas. El valor p de una prueba F conjunta de todas las covariables listadas en una regresión del estado de tratamiento sobre esas covariables, con efectos fijos de estrato de aleatorización y errores estándar agrupados por escuela, es < 0.001 . * significativo al 10%; ** significativo al 5%; *** significativo al 1%.

Tabla A.3: Pruebas de balance de características y habilidades docentes (línea de base)

Variable	(1) Control	(2) Tratamiento	(3) Col.(2)-(1)
<i>A. Características de contexto</i>			
Mujer	0.12 [0.32]	0.15 [0.36]	0.03 (0.04)
Edad	41.7 [8.67]	41.8 [7.64]	0.04 (0.93)
N (docentes)	156	155	311
<i>B. Educación y experiencia</i>			
Título en educación	0.07 [0.26]	0.09 [0.29]	0.02 (0.03)
Título en matemática	0.31 [0.47]	0.28 [0.45]	-0.03 (0.05)
Tiene título de maestría	0.31 [0.47]	0.36 [0.48]	0.05 (0.05)
Tiene formación inicial	0.28 [0.45]	0.18 [0.39]	-0.10** (0.05)
Tiene formación en servicio	0.47 [0.50]	0.52 [0.50]	0.06 (0.06)
Experiencia docente	14.5 [9.16]	15.3 [9.13]	0.85 (1.03)
Experiencia en esta escuela	12.9 [9.27]	13.9 [8.75]	1.00 (1.03)
N (docentes)	156	155	311
<i>C. Habilidades socioemocionales</i>			
Autoeficacia (est.)	0.00 [1.00]	0.11 [0.90]	0.11 (0.11)
Satisfacción laboral (est.)	0.00 [1.00]	0.01 [0.93]	0.01 (0.11)
Locus de control (est.)	0.00 [1.00]	0.10 [1.04]	0.10 (0.11)
Determinación (est.)	0.00 [1.00]	0.01 [1.01]	0.00 (0.11)
Mentalidad de crecimiento (est.)	0.00 [1.00]	0.06 [1.14]	0.06 (0.12)
N (docentes)	156	155	311

Notas: La tabla muestra las medias y los DE de los docentes de los grupos de control (columna 1) y de tratamiento (columna 2). También prueba las diferencias entre grupos, teniendo en cuenta los efectos fijos de estrato de aleatorización (columna 3). El panel A muestra los resultados de las características de contexto, el panel B los de educación y experiencia y el panel C los de las habilidades socioemocionales. El valor p de una prueba F conjunta de todas las covariables listadas en una regresión del estado de tratamiento sobre esas covariables, con efectos fijos de estrato de aleatorización y errores estándar agrupados por escuela, es 0.499. * significativo al 10%; ** significativo al 5%; *** significativo al 1%.

Tabla A.4: Atrición en las evaluaciones de estudiantes (línea de cierre)

	(1) Con atrición
<i>A. Tratamiento</i>	
Tratamiento	0.02 (0.02)
N (estudiantes)	5940
Media de control	0.23
<i>B. Tratamiento y línea de base</i>	
Tratamiento	0.01 (0.15)
Mujer	-0.06*** (0.02)
Edad	0.05*** (0.01)
Puntaje compuesto (est.)	-0.03*** (0.01)
Tratamiento $\times$ Mujer	-0.01 (0.03)
Tratamiento $\times$ Edad	0.00 (0.01)
Tratamiento $\times$ Puntaje compuesto	0.01 (0.01)
N (estudiantes)	5921
Razón F (Interacciones)	0.75
Valor p	0.53

Notas: La tabla muestra las estimaciones de regresiones que predicen el estado de seguimiento en las evaluaciones de cierre. El seguimiento se define como tener un puntaje observado en la línea de cierre. La muestra incluye a los presentes en la línea de base. El panel A regresa la atrición sobre el tratamiento, y el panel B sobre el tratamiento interactuado con las características de línea de base. Ambos paneles incluyen efectos fijos de estrato de aleatorización. Los errores estándar (agrupados por escuela) aparecen entre paréntesis. Los valores F y p corresponden a una prueba conjunta de todos los términos de interacción. * significativo al 10 %; ** significativo al 5 %; *** significativo al 1 %.

Tabla A.5: Estimación docente del rendimiento estudiantil en tareas concretas (línea de base)

	(1) Real	(2) Estimado	(3) Diferencia
<i>A. Porcentaje de estudiantes competentes, por tema</i>			
<u>Número</u>	5.40	41.8	36.5*** (1.40)
<u>Álgebra</u>	18.1	52.1	34.0*** (1.60)
<u>Medición</u>	0.30	36.0	35.6*** (1.30)
<u>Geometría</u>	2.70	40.8	38.1*** (1.50)
<u>Datos</u>	0.90	45.7	44.8*** (1.50)
N (docentes)	311	311	311
<i>B. Porcentaje de estudiantes que responde correctamente ítems específicos</i>			
<u>Operaciones</u>	56.9	39.6	-17.3*** (2.10)
<u>Geometría</u>	30.6	42.7	12.2*** (1.90)
<u>Datos</u>	48.3	50.8	2.60 (2.00)
N (docentes)	311	311	311
<i>C. Probabilidad de que estudiantes específicos respondan correctamente ítems específicos</i>			
<u>Operaciones</u>	0.82	0.90	0.08*** (0.02)
<u>Geometría</u>	0.24	0.54	0.30*** (0.02)
N (respuestas estudiante-docente-ítem)	3,063	3,063	3,063

Notas: La tabla muestra el desempeño real y el estimado de los estudiantes en tareas concretas de matemática. El panel A muestra el porcentaje real y estimado de estudiantes “competentes” (es decir, aquellos que pueden responder correctamente la mayoría de las preguntas) en cada tema, a nivel de docente. El panel B muestra el porcentaje real y estimado de estudiantes que pueden responder correctamente tres ítems, a nivel de docente: en operaciones, “¿Cuál es el signo que falta en $72 - 6 + 15 = 27$?”; en medición y geometría, “Nombre un triángulo con dos lados iguales”; y en datos, “Halle la moda de 8, 12, 11, 4, 7, 12, 62.” El panel C muestra la probabilidad real y estimada de que estudiantes específicos puedan responder correctamente dos ítems, a nivel de estudiante-docente-ítem: en operaciones, “¿Cuánto es 15×4 ?”; y en medición y geometría, “Una caja rectangular tiene 6 cm de largo, 4 cm de ancho y 2 cm de espesor. ¿Cuál es el volumen de la caja?” Las diferencias equivalen a lo estimado menos lo real y provienen de regresiones de los valores reales y estimados sobre un indicador de los estimados, con efectos fijos de estrato de aleatorización. Los errores estándar agrupados a nivel de escuela aparecen entre paréntesis. La tabla incluye ambos grupos experimentales en la línea de base. * significativo al 10 %; ** significativo al 5 %; *** significativo al 1 %.

Tabla A.6: Observaciones de aula Stallings (línea de cierre)

	(1) Control	(2) Tratamiento	(3) Col.(2)-(1)
<i>A. Asignación del tiempo de clase (%)</i>			
En tarea	93.6 [10.3]	93.3 [10.9]	-0.24 (1.17)
Gestión de la clase	1.76 [4.59]	1.84 [5.46]	0.08 (0.54)
Fuera de tarea	4.65 [9.55]	4.81 [9.89]	0.16 (1.07)
<i>B. Tiempo de clase en tarea, por actividad (%)</i>			
Lectura en voz alta	9.38 [19.9]	9.30 [18.0]	-0.08 (1.92)
Explicación	22.2 [21.4]	22.2 [19.6]	0.00 (2.09)
Demostración	10.5 [19.4]	11.7 [17.5]	1.20 (1.82)
Discusión	22.8 [22.9]	23.6 [24.3]	0.80 (2.48)
Práctica y ejercitación	11.1 [17.2]	9.38 [15.2]	-1.68 (1.76)
Consigna	6.89 [11.5]	7.13 [12.1]	0.24 (1.30)
Copia	3.13 [7.99]	3.37 [8.91]	0.24 (0.92)
Instrucciones	7.61 [12.4]	6.65 [11.7]	-0.96 (1.28)
<i>C. Participación de los estudiantes (% del tiempo de clase)</i>			
Ningún estudiante	5.45 [11.0]	6.09 [11.3]	0.64 (1.20)
1 estudiante	6.81 [12.6]	7.53 [13.3]	0.72 (1.41)
2-10 estudiantes	7.29 [12.8]	8.89 [14.6]	1.60 (1.51)
10+ estudiantes	11.8 [21.5]	11.5 [21.1]	-0.24 (2.21)
Todos los estudiantes	68.7 [26.9]	65.9 [28.5]	-2.72 (3.00)
N (escuelas)	156	156	312

Notas: La tabla presenta los resultados de las observaciones de aula de estilo Stallings de la línea de cierre. El panel A muestra el porcentaje del tiempo de clase asignado a actividades de tarea, gestión de la clase y actividades ajenas a la tarea. El panel B descompone las actividades de tarea como porcentaje del tiempo de clase. El panel C muestra la participación de los estudiantes como porcentaje del tiempo de clase. Cada escuela contribuye con una observación. Las columnas 1 y 2 reportan las medias y los DE (entre corchetes) de cada resultado en los grupos de control y de tratamiento, respectivamente; la columna 3 reporta la diferencia, estimada por una regresión sobre el indicador de tratamiento con efectos fijos de estrato. Los errores estándar robustos a heterocedasticidad están entre paréntesis. * significativo al 10 %; ** significativo al 5 %; *** significativo al 1 %.

Tabla A.7: Interacciones entre estudiantes y docentes según el rendimiento de línea de base en el grupo de control (línea de cierre)

	(1) Todos los estudiantes	(2) Bajo rendimiento	(3) Rendimiento medio	(4) Alto rendimiento	(5) p (2)=(3)	(6) p (2)=(4)
<i>A. Global</i>						
Cantidad de interacciones	0.24 (0.01)	0.15 (0.02)	0.22 (0.02)	0.36 (0.02)	< 0.01	< 0.01
Proporción con 0 interacciones	0.80 (0.01)	0.87 (0.01)	0.80 (0.01)	0.71 (0.02)	< 0.01	< 0.01
<i>B. Según quién inició</i>						
Docente	0.16 (0.01)	0.11 (0.01)	0.15 (0.01)	0.24 (0.02)	0.01	< 0.01
Estudiante	0.08 (0.01)	0.04 (0.01)	0.07 (0.01)	0.12 (0.01)	0.01	< 0.01
<i>C. Por tipo</i>						
Afirmación	0.08 (0.01)	0.06 (0.01)	0.06 (0.01)	0.13 (0.01)	0.57	< 0.01
Pregunta	0.16 (0.01)	0.10 (0.01)	0.16 (0.02)	0.23 (0.02)	< 0.01	< 0.01
<i>D. Por dirección</i>						
Unidireccional	0.03 (0.01)	0.02 (0.01)	0.02 (0.01)	0.05 (0.01)	0.59	0.01
Bidireccional	0.16 (0.01)	0.11 (0.01)	0.16 (0.01)	0.23 (0.02)	< 0.01	< 0.01
Multidireccional	0.05 (0.01)	0.03 (0.01)	0.04 (0.01)	0.08 (0.01)	0.20	< 0.01
<i>E. Por contenido</i>						
Académica	0.22 (0.01)	0.14 (0.02)	0.21 (0.02)	0.34 (0.02)	< 0.01	< 0.01
Logística	0.01 (0.00)	0.01 (0.00)	0.01 (0.00)	0.01 (0.00)	0.11	0.69
Social	0.01 (0.00)	0.00 (0.00)	0.00 (0.00)	0.01 (0.00)	0.43	0.12
<i>F. Por duración</i>						
< 5 segundos	0.04 (0.01)	0.02 (0.01)	0.04 (0.01)	0.06 (0.01)	0.06	< 0.01
6-10 segundos	0.08 (0.01)	0.05 (0.01)	0.08 (0.01)	0.12 (0.01)	0.01	< 0.01
> 10 segundos	0.12 (0.01)	0.08 (0.01)	0.11 (0.01)	0.17 (0.02)	0.12	< 0.01
N (estudiantes)	2,964	1,041	1,030	893		

Notas: La tabla muestra las medias ajustadas de la cantidad de interacciones entre estudiantes y docentes por estudiante por clase observadas en la línea de cierre en las escuelas de control, en forma global (panel A), según quién inició la interacción (panel B), por tipo de interacción (panel C), por dirección de la interacción (panel D), por contenido (panel E) y por duración de la interacción (panel F). Los resultados son conteos observados por estudiante durante una única clase de observación de aula. La columna 1 muestra los resultados de todos los estudiantes de control, y las columnas 2-4 los de los estudiantes de control de los terciles inferior, medio y superior de rendimiento en matemática de línea de base dentro del aula en las evaluaciones de impacto, respectivamente. Las columnas 5 y 6 reportan los valores p de las pruebas de que las medias ajustadas de las columnas 2 y 3, y de las columnas 2 y 4, son iguales. Todas las especificaciones restringen la muestra a los estudiantes del grupo de control e incluyen efectos fijos de estrato de aleatorización. Los errores estándar agrupados a nivel de escuela aparecen entre paréntesis.

Tabla A.8: Interacciones entre estudiantes y docentes según las características de los docentes en el grupo de control (línea de cierre)

	(1)	(2)	(3)	(4)	(5)	(6)
	Sí/ ≥ mediana		No/ < mediana		Col. (1)-(3)	
	Media	DE	Media	DE	Dif.	EE
<i>A. Demografía</i>						
Mujer	7.56	[4.37]	8.67	[5.40]	-0.69	(0.64)
Edad	8.73	[5.15]	8.33	[5.48]	1.02*	(0.53)
<i>B. Educación</i>						
Título en educación	7.18	[3.66]	8.65	[5.39]	-3.41**	(1.68)
Título en matemática	8.61	[4.21]	8.51	[5.74]	0.83	(0.60)
Tiene título de maestría	7.20	[3.35]	9.16	[5.89]	-0.59	(0.55)
Tiene formación inicial	7.86	[4.19]	8.81	[5.66]	0.30	(0.63)
Tiene formación en servicio	8.63	[5.98]	8.47	[4.64]	0.25	(0.61)
<i>C. Experiencia</i>						
Experiencia docente	8.93	[5.25]	8.20	[5.33]	0.97*	(0.54)
Experiencia en esta escuela	8.89	[5.31]	8.24	[5.29]	1.00*	(0.54)
<i>D. Estimaciones de los docentes</i>						
Brecha en las estimaciones	8.61	[5.85]	8.43	[4.85]	-0.03	(0.59)
<i>E. Habilidades socioemocionales</i>						
Índice (est.)	8.32	[5.22]	8.80	[5.40]	-0.42	(0.59)
Autoeficacia	7.38	[3.96]	9.78	[6.19]	-1.17**	(0.53)
Satisfacción laboral	8.52	[5.47]	8.60	[4.92]	-0.22	(0.62)
Locus de control	8.70	[5.28]	7.90	[5.38]	1.06	(0.65)
Determinación	8.76	[5.27]	8.16	[5.35]	0.67	(0.61)
Mentalidad de crecimiento	9.28	[6.03]	7.78	[4.29]	0.18	(0.52)
N (docentes)	156					

Notas: La tabla muestra la cantidad de interacciones entre estudiantes y docentes por clase observadas en la línea de cierre en las escuelas de control, según las características de línea de base de los docentes. Para las características binarias, las columnas 1 y 2 muestran a los docentes con la característica y las columnas 3 y 4 a los docentes sin ella. Para las características continuas, las columnas 1 y 2 muestran a los docentes en o por encima de la mediana y las columnas 3 y 4 a los docentes por debajo de la mediana. El índice es el primer componente principal estandarizado de autoeficacia, satisfacción laboral, locus de control, determinación y mentalidad de crecimiento. Las columnas 1-4 reportan medias y DE entre corchetes. Las columnas 5 y 6 reportan las diferencias estimadas a partir de regresiones con efectos fijos de estrato de aleatorización. Los errores estándar agrupados a nivel de escuela aparecen entre paréntesis. * significativo al 10%; ** significativo al 5%; *** significativo al 1%.

Tabla A.9: Interacciones entre estudiantes y docentes según las características de los estudiantes en el grupo de control (línea de cierre)

	(1)	(2)	(3)	(4)	(5)	(6)
	Sí/ ≥ mediana		No/ < mediana		Col. (1)-(3)	
	Media	DE	Media	DE	Dif.	EE
<i>A. Características de contexto</i>						
Mujer	0.23	[0.53]	0.25	[0.59]	0.02	(0.02)
Edad	0.22	[0.54]	0.29	[0.59]	-0.06***	(0.02)
Habla bangla en el hogar	0.21	[0.49]	0.68	[1.09]	0.20	(0.24)
Madre completó la secundaria	0.26	[0.52]	0.22	[0.58]	0.09***	(0.02)
Activos del hogar (est.)	0.28	[0.62]	0.19	[0.46]	0.10***	(0.02)
<i>B. Habilidades socioemocionales</i>						
Índice (est.)	0.28	[0.59]	0.19	[0.52]	0.09***	(0.02)
Mentalidad de crecimiento (est.)	0.28	[0.58]	0.18	[0.52]	0.10***	(0.02)
Autogestión (est.)	0.26	[0.58]	0.21	[0.52]	0.05**	(0.02)
Autoeficacia (est.)	0.28	[0.59]	0.19	[0.50]	0.09***	(0.02)
Orientación a metas de tarea (est.)	0.24	[0.53]	0.24	[0.62]	0.01	(0.02)
Motivación por dominar tareas (est.)	0.26	[0.55]	0.21	[0.57]	0.06***	(0.02)
Miedo al fracaso (est.)	0.24	[0.56]	0.23	[0.56]	0.00	(0.03)
Determinación (est.)	0.22	[0.51]	0.26	[0.61]	-0.04**	(0.02)
<i>C. Habilidades académicas</i>						
Índice (est.)	0.31	[0.64]	0.16	[0.45]	0.15***	(0.02)
Matemática (TRI, est.)	0.31	[0.63]	0.16	[0.46]	0.15***	(0.02)
Lectura (TRI, est.)	0.29	[0.62]	0.19	[0.48]	0.10***	(0.02)
Inteligencia fluida (est.)	0.27	[0.59]	0.19	[0.50]	0.08***	(0.02)
N (estudiantes)	2,964					

Notas: La tabla muestra la cantidad de interacciones entre estudiantes y docentes por estudiante durante la observación de aula de la línea de cierre en las escuelas de control, según las características de línea de base de los estudiantes. Para las características binarias, las columnas 1 y 2 muestran a los estudiantes con la característica y las columnas 3 y 4 a los estudiantes sin ella. Para las características continuas, las columnas 1 y 2 muestran a los estudiantes en o por encima de la mediana del aula y las columnas 3 y 4 a los estudiantes por debajo de la mediana del aula. El índice socioemocional es el primer componente principal estandarizado de mentalidad de crecimiento, autogestión, autoeficacia, miedo al fracaso (codificado en sentido inverso), orientación a metas, motivación por dominar tareas y determinación. El índice académico es el primer componente principal estandarizado de rendimiento en matemática, rendimiento en lectura e inteligencia fluida. Las columnas 1-4 reportan medias y DE entre corchetes. Las columnas 5 y 6 reportan las diferencias estimadas a partir de regresiones con efectos fijos de estrato de aleatorización. Los errores estándar agrupados a nivel de escuela aparecen entre paréntesis. * significativo al 10%; ** significativo al 5%; *** significativo al 1%.

Tabla A.10: Interacciones entre estudiantes y docentes y precisión de las estimaciones de los docentes en el grupo de control (línea de cierre)

	(1) Todos los estudiantes	(2) Bajo rendimiento	(3) Rendimiento medio	(4) Alto rendimiento	(5) p (2)=(3)	(6) p (2)=(4)
Brecha de puntaje (DE)	0.17* (0.09)	0.03 (0.10)	0.29** (0.14)	0.21** (0.10)	0.08	0.15
Brecha absoluta de puntaje (DE)	-0.18** (0.09)	-0.11 (0.09)	-0.20 (0.14)	-0.14* (0.08)	0.46	0.75
Brecha de ordenamiento (percentiles)	0.03* (0.02)	0.01 (0.03)	0.08*** (0.03)	0.03 (0.02)	0.10	0.55
Brecha absoluta de ordenamiento (percentiles)	-0.02* (0.01)	-0.01 (0.02)	0.00 (0.02)	-0.03** (0.01)	0.58	0.46
El docente sobreestima	0.01 (0.02)	-0.01 (0.02)	0.02 (0.03)	0.04 (0.03)	0.55	0.15
N (díadas estudiante-docente)	2,189	722	759	708		

Notas: La tabla muestra los coeficientes de regresiones de los resultados de las estimaciones de los docentes sobre la cantidad de interacciones observadas entre el docente y el estudiante durante la observación de aula de la línea de cierre. La columna 1 usa a todos los estudiantes del grupo de control; las columnas 2-4 usan a los estudiantes del grupo de control de los terciles inferior, medio y superior de rendimiento en matemática de línea de base dentro del aula en las evaluaciones de impacto, respectivamente. Las columnas 5 y 6 reportan los valores p de las pruebas de que los coeficientes de las columnas 2 y 3, y de las columnas 2 y 4, son iguales. Las brechas de puntaje equivalen a los puntajes estimados menos los reales, escaladas por el DE dentro del aula de los puntajes reales; las brechas de ordenamiento equivalen a los rangos percentilares estimados menos los reales dentro del aula; las brechas absolutas son los valores absolutos de esas medidas. Todas las especificaciones restringen la muestra a los estudiantes del grupo de control e incluyen efectos fijos de escuela. Los errores estándar agrupados a nivel de escuela aparecen entre paréntesis. * significativo al 10%; ** significativo al 5%; *** significativo al 1%.

Tabla A.11: Impacto sobre el rendimiento en matemática según el estado de señalamiento (línea de cierre)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Señalados		No señalados		Diferencia		Valor p de igualdad	
	Según informe	Según evaluación	Según informe	Según evaluación	Según informe	Según evaluación	Según informe	Según evaluación
Matemática global	0.21*** (0.08)	0.24*** (0.08)	0.23*** (0.08)	0.21*** (0.07)	-0.02 (0.06)	0.03 (0.05)	0.76	0.57
Número	0.19* (0.10)	0.26*** (0.08)	0.25*** (0.06)	0.23*** (0.06)	-0.05 (0.09)	0.03 (0.06)	0.54	0.60
Medición y geometría	0.14 (0.09)	0.21** (0.08)	0.17** (0.07)	0.15** (0.07)	-0.02 (0.07)	0.06 (0.06)	0.78	0.32
Datos	0.12 (0.10)	0.17** (0.08)	0.16** (0.07)	0.15** (0.07)	-0.04 (0.08)	0.02 (0.06)	0.65	0.80
N (estudiantes)	4,543							

Notas: La tabla muestra los impactos sobre el rendimiento en matemática de cierre según el estado de señalamiento. Dentro de cada par de columnas, las columnas basadas en los informes usan los señalamientos observados en los informes para los estudiantes de tratamiento y los tres estudiantes de más bajo desempeño dentro de cada clase en cada tema de las evaluaciones de impacto de línea de base para los estudiantes de control; las columnas basadas en las evaluaciones de impacto usan esta regla de los tres de más bajo desempeño en ambos grupos. Las columnas 1-2 reportan los efectos para los estudiantes señalados; las columnas 3-4 reportan los efectos para los estudiantes no señalados; las columnas 5-6 reportan las diferencias entre estos efectos; y las columnas 7-8 reportan los valores p de las pruebas de su igualdad. Los puntajes escalados por TRI están estandarizados para tener una media de cero y un DE de uno en el grupo de control global. Todas las especificaciones incluyen efectos fijos de estrato de aleatorización y covariables de línea de base seleccionadas mediante LASSO. Los errores estándar agrupados a nivel de escuela aparecen entre paréntesis. * significativo al 10%; ** significativo al 5%; *** significativo al 1%.

Tabla A.12: Impacto sobre el rendimiento en lectura según el estado de señalamiento en matemática (línea de cierre)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Señalados		No señalados		Diferencia		Valor p de igualdad	
	Según informe	Según evaluación	Según informe	Según evaluación	Según informe	Según evaluación	Según informe	Según evaluación
Lectura global	0.15*	0.13	0.20**	0.20**	-0.05	-0.07	0.45	0.13
	(0.09)	(0.08)	(0.08)	(0.08)	(0.07)	(0.05)		
Palabras	0.12	0.10	0.18**	0.18***	-0.06	-0.08	0.35	0.12
	(0.08)	(0.08)	(0.07)	(0.07)	(0.07)	(0.05)		
Oraciones	0.13	0.12	0.14*	0.15**	-0.01	-0.03	0.87	0.58
	(0.08)	(0.08)	(0.07)	(0.07)	(0.07)	(0.06)		
Párrafos	0.04	0.03	0.10	0.10	-0.06	-0.08	0.41	0.19
	(0.08)	(0.08)	(0.08)	(0.08)	(0.07)	(0.06)		
Relato	0.16*	0.07	0.12	0.13	0.04	-0.06	0.58	0.35
	(0.09)	(0.07)	(0.08)	(0.09)	(0.07)	(0.06)		
Vocabulario	0.12	0.11	0.18**	0.18**	-0.06	-0.08	0.33	0.14
	(0.09)	(0.08)	(0.07)	(0.07)	(0.07)	(0.05)		
Comprensión	0.15*	0.10	0.17**	0.17**	-0.02	-0.07	0.83	0.21
	(0.09)	(0.08)	(0.08)	(0.08)	(0.07)	(0.06)		
N (estudiantes)					4,543			

Notas: La tabla muestra los impactos sobre el rendimiento en lectura de cierre según el estado de señalamiento en matemática. Dentro de cada par de columnas, las columnas basadas en los informes usan los señalamientos observados en los informes para los estudiantes de tratamiento y los tres estudiantes de más bajo desempeño dentro de cada clase en cada tema de matemática de las evaluaciones de impacto de línea de base para los estudiantes de control; las columnas basadas en las evaluaciones de impacto usan esta regla de los tres de más bajo desempeño en ambos grupos. Las columnas 1-2 reportan los efectos para los estudiantes señalados; las columnas 3-4 reportan los efectos para los estudiantes no señalados; las columnas 5-6 reportan las diferencias entre estos efectos; y las columnas 7-8 reportan los valores p de las pruebas de su igualdad. Los puntajes escalados por TRI están estandarizados para tener una media de cero y un DE de uno en el grupo de control global. Todas las especificaciones incluyen efectos fijos de estrato de aleatorización y covariables de línea de base seleccionadas mediante LASSO. Los errores estándar agrupados a nivel de escuela aparecen entre paréntesis. * significativo al 10%; ** significativo al 5%; *** significativo al 1%.

Tabla A.13: Impacto sobre las estimaciones específicas por tema de los docentes sobre los puntajes de los estudiantes (línea de cierre)

	(1) Todos los estudiantes	(2) Bajo rendimiento	(3) Rendimiento medio	(4) Alto rendimiento	(5) p (2)=(3)	(6) p (2)=(4)
<i>A. Brecha en el % de estudiantes competentes, por tema (p.p.)</i>						
<u>Número</u>	4.30 (2.92)	3.92 (2.96)	4.24 (2.97)	4.27 (2.96)	0.76	0.70
Media de control	41.6	42.1	41.5	41.3		
<u>Medición y geometría</u>	0.94 (2.51)	0.43 (2.55)	1.75 (2.57)	0.99 (2.52)	0.10	0.48
Media de control	46.5	47.1	46.3	46.0		
<u>Datos</u>	3.69 (2.24)	3.88* (2.27)	4.13* (2.32)	3.07 (2.25)	0.72	0.26
Media de control	65.2	65.6	64.5	65.5		
<i>B. Brecha en el % de estudiantes que responde correctamente ítems específicos, por tema (p.p.)</i>						
<u>Operaciones</u>	-1.01 (3.24)	-1.27 (3.27)	-0.52 (3.40)	-1.40 (3.25)	0.53	0.91
Media de control	-5.50	-5.23	-5.56	-5.69		
<u>Geometría</u>	1.23 (3.11)	0.27 (3.22)	2.00 (3.14)	1.39 (3.15)	0.14	0.30
Media de control	-1.45	-0.29	-2.14	-1.91		
<u>Datos</u>	-2.81 (3.42)	-3.86 (3.49)	-2.05 (3.48)	-2.79 (3.36)	0.10	0.37
Media de control	15.6	16.5	14.8	15.5		
N (díadas estudiante-docente)	4,314	1,423	1,503	1,388		

Notas: Tabla del apéndice correspondiente a la Tabla 5. La tabla muestra el impacto de la intervención sobre la brecha entre el % de estudiantes competentes estimado y el real (panel A) y el % que responde correctamente ítems específicos (panel B), por tema. Los coeficientes negativos (positivos) indican subestimación (sobreestimación). Las cols. 2-4 son terciles dentro del aula de rendimiento en matemática de línea de base en las evaluaciones de impacto; las cols. 5-6 dan los valores p de las cols. 2 vs. 3 y 2 vs. 4. Todas las especificaciones incluyen EF de estrato y covariables de línea de base seleccionadas por LASSO. Los EE agrupados a nivel de escuela entre paréntesis. * 10%; ** 5%; *** 1%.

Tabla A.14: Impacto sobre los temas enseñados durante las clases observadas (línea de cierre)

	(1) Control	(2) Tratamiento	(3) Col.(2)-(1)
Número	0.73 [0.45]	0.86 [0.35]	0.14 (0.09)
Medición y geometría	0.14 [0.35]	0.14 [0.35]	-0.01 (0.08)
Datos	0.18 [0.39]	0.04 [0.20]	-0.15** (0.07)
N (escuelas)	44	49	93

Notas: La tabla muestra el impacto de la intervención sobre si la clase observada cubrió cada tema, medido en la línea de cierre en una submuestra aleatoria de escuelas. Cada escuela contribuye con una observación. Las columnas 1 y 2 reportan las medias y los DE (entre corchetes) de cada resultado en los grupos de control y de tratamiento, respectivamente. La columna 3 reporta la diferencia, estimada por una regresión sobre la variable indicadora de tratamiento con efectos fijos de estrato de aleatorización; los errores estándar robustos a heterocedasticidad están entre paréntesis. * significativo al 10 %; ** significativo al 5 %; *** significativo al 1 %.

Tabla A.15: Impacto sobre la asistencia estudiantil y docente (línea intermedia y de cierre)

	(1) Todos los estudiantes	(2) Bajo rendimiento	(3) Rendimiento medio	(4) Alto rendimiento	(5) p (2)=(3)	(6) p (2)=(4)
<i>A. Asistencia de los estudiantes</i>						
<u>Registros escolares</u>						
Línea intermedia	-0.01 (0.02)	-0.01 (0.02)	-0.02 (0.02)	-0.01 (0.02)	0.10	0.74
Media de control	0.70	0.68	0.71	0.73		
N (estudiantes)	5,935	2,084	2,063	1,787		
Línea de cierre	0.02 (0.01)	0.04** (0.02)	-0.01 (0.02)	0.02 (0.02)	< 0.01	0.22
Media de control	0.71	0.66	0.72	0.74		
N (estudiantes)	5,917	2,077	2,059	1,780		
<u>Asistencia observada</u>						
Línea intermedia	0.01 (0.02)					
Media de control	0.62					
N (escuelas)	311					
Línea de cierre	0.02 (0.02)					
Media de control	0.61					
N (escuelas)	310					
<i>B. Asistencia de los docentes</i>						
<u>Registros escolares</u>						
Línea intermedia	0.01* (0.01)					
Media de control	0.93					
N (docentes)	304					
Línea de cierre	0.00 (0.01)					
Media de control	0.93					
N (docentes)	306					

Notas: La tabla muestra los impactos sobre la asistencia de los estudiantes (panel A) y la asistencia de los docentes (panel B) en la línea intermedia y en la de cierre. La asistencia según los registros escolares se calcula como los días presentes divididos por los días laborables. La asistencia observada se calcula como la cantidad de estudiantes presentes el día de la encuesta dividida por la matrícula. La columna 1 reporta los efectos para todos los estudiantes y los efectos agrupados de los docentes. Las columnas 2-4 reportan los efectos sobre la asistencia de los estudiantes basada en registros para los estudiantes de los terciles inferior, medio y superior de rendimiento en matemática de línea de base dentro del aula en las evaluaciones de impacto, respectivamente. Las columnas 5 y 6 reportan los valores p de las pruebas de igualdad de los efectos de tratamiento de las columnas 2 y 3, y de las columnas 2 y 4, respectivamente. Las medias de control y los tamaños de muestra aparecen debajo de cada conjunto de estimaciones. Todas las especificaciones incluyen efectos fijos de estrato de aleatorización; las especificaciones de asistencia observada también controlan por la asistencia observada de línea de base. Los errores estándar agrupados a nivel de escuela aparecen entre paréntesis. * significativo al 10 %; ** significativo al 5 %; *** significativo al 1 %.

Apéndice B Componentes de la intervención

La intervención constó de tres componentes, que se describen a continuación.

B.1 Evaluaciones de la intervención

Los estudiantes de las escuelas de tratamiento completaron dos evaluaciones breves de matemática de 10 ítems durante el año escolar 2022. La primera se administró en mayo y la segunda en agosto (Figura A.3).³² Decidimos la distribución de los ítems entre temas (número, medición y geometría, y datos) y habilidades (conocer, aplicar y razonar) con base en el marco de evaluación del Estudio de las Tendencias en Matemática y Ciencias (TIMSS, por su sigla en inglés), una evaluación internacional de matemática y ciencias para 4.º y 8.º grado administrada por la Asociación Internacional para la Evaluación del Rendimiento Educativo (IEA, 2021). Luego ajustamos el marco con base en el currículo nacional bangladésí (NCTB, 2023). El marco de las evaluaciones de la intervención se muestra en las Tablas B.1 y B.2.

B.2 Informes para los docentes

Después de cada evaluación, los docentes recibieron un informe de aprendizaje breve, de cinco páginas (Figuras B.1-B.5). Cada informe resumía el propósito del ejercicio, el desempeño global de la clase, el desempeño de los estudiantes en cada ítem, el desempeño por tema, los estudiantes con mayor necesidad de apoyo en cada tema, las operaciones que parecían necesitar refuerzo y sugerencias concretas para usar los resultados en la enseñanza. Los informes usaban los colores del semáforo para los resultados a nivel de ítem: rojo cuando no más del 32 % de los estudiantes respondía correctamente, amarillo cuando lo hacía entre el 33 y el 65 %, y verde cuando lo hacía al menos el 66 %. Los informes privilegiaban las descripciones narrativas de los resultados en vez de gráficos, y su formato se refinó en tres rondas de comentarios con docentes bangladésíes para hacer la información más clara, más fácil de interpretar y más útil.

B.3 Guía de análisis de errores

Los docentes recibieron también una guía de análisis de errores que vinculaba las respuestas incorrectas frecuentes con posibles concepciones erróneas. Para cada ítem, la guía identificaba la respuesta correcta, describía cómo obtenerla e interpretaba las concepciones erróneas más probablemente subyacentes a cada respuesta incorrecta (Figuras B.6 y B.7).

³²El año escolar en Bangladesh había comenzado en febrero después de los cierres relacionados con el COVID; marzo y abril se utilizaron para pilotear las pruebas y los instrumentos de recolección de datos.

Figura B.1: Plantilla del informe de aprendizaje, página 1: presentación general y propósito

Escuela Secundaria Pública

EIIN N.º 99999

INFORME SOBRE EL APRENDIZAJE DE LOS ESTUDIANTES

¿Por qué recibe este informe?

El Gobierno de Bangladesh y el Banco Mundial se han asociado para brindar a las escuelas datos sobre el aprendizaje de los estudiantes, a fin de ayudarlo a identificar a los estudiantes que necesitan apoyo adicional.

¿Qué información incluye?

El informe resume el desempeño de sus estudiantes en habilidades matemáticas fundamentales, con base en las evaluaciones realizadas en agosto de 2022 en su escuela. También compara su desempeño con el de los estudiantes de escuelas similares de su distrito y muestra las mejoras de sus estudiantes desde junio de 2022.

¿Por qué es importante?

Los estudiantes necesitan dominar las habilidades matemáticas fundamentales para comprender el currículo de 6.º grado. Aunque usted puede no haber sido responsable de enseñarles estas habilidades, es importante que atienda cualquier brecha en sus conocimientos para ayudarlos a comprender mejor sus clases.

¿Cómo debería usarlo?

- Identifique las *áreas* (por ejemplo, números y operaciones) y los *temas* (por ejemplo, resta) en los que sus estudiantes tienen el desempeño más bajo;
- Prepare una o más clases que cubran las áreas y los temas que necesitan refuerzo, verificando durante la clase si los estudiantes comprenden el material;
- Contacte a los estudiantes con el desempeño más bajo en estas áreas y temas según este informe, revise sus tareas y pregúnteles si necesitan ayuda adicional;
- Haga un seguimiento del progreso de los estudiantes de bajo desempeño usando el informe de la ronda anterior que recibió en junio de 2022

¿Cómo sabrá si sus estudiantes mejoraron?

Visitamos su escuela en junio de 2022 y le entregamos un informe similar; compararlo con este le dará una idea del progreso de sus estudiantes.

Notas: La figura muestra la primera página del informe de aprendizaje de cinco páginas entregado a los docentes de tratamiento después de cada evaluación de la intervención. La página presenta el informe, enuncia su propósito, resume cómo fue elaborado y explica cómo los docentes deberían interpretar la información que sigue.

Figura B.2: Plantilla del informe de aprendizaje, página 2: desempeño global y por ítem

¿Cómo se desempeñaron sus estudiantes en comparación con otros de su distrito?

- El estudiante típico de su clase respondió correctamente **6 de 10** (60%) de las preguntas, en comparación con **6 de 10** (60%) preguntas en su distrito.
- En su clase, el **31%** de sus estudiantes tiene **bajo desempeño**, el **35%** tiene **desempeño medio** y el **11%** tiene **alto desempeño**, en comparación con el **37%, 35% y 27%** en su distrito.
- Los estudiantes de bajo desempeño de su clase respondieron correctamente **4 de 10** (40%) preguntas, en comparación con **4 de 10** (40%) en su distrito. Estos son los estudiantes que más necesitan su apoyo.
- El área en la que los estudiantes de su clase tuvieron el desempeño más bajo fue **medición y geometría**.
- Los temas en los que los estudiantes de su clase más necesitan ayuda fueron **suma de números y conversión de unidades de medida y aplicación de propiedades geométricas para hallar un ángulo faltante**.

¿Cómo se desempeñaron sus estudiantes en cada pregunta?

Área de contenido	Número de pregunta en la prueba, % de respuestas correctas en esta pregunta				
Números y operaciones	P1 60%	P2 85%	P3 88%	P4 84%	P5 95%
Geometría y medición	P7 3%	P8 39%	P9 12%		
Datos	P6 65%	P10 39%			

- P1: Suma de un número entero y un número decimal.
P2: Multiplicación de un número de un dígito por un número de dos dígitos.
P3: Problema con enunciado que involucra la resta de un número de un dígito a otro.
P4: División de un número de dos dígitos por un número de un dígito.
P5: Problema con enunciado que involucra multiplicación cruzada.
P6: Interpretación de un calendario para hacer extrapolaciones.
P7: Suma de números y conversión de unidades de medida.
P8: Cálculo del volumen de una figura tridimensional.
P9: Aplicación de propiedades geométricas para hallar un ángulo faltante.
P10: Interpretación de tablas con datos para calcular una fracción.

Notas: La figura muestra la segunda página del informe de aprendizaje. Reporta el desempeño global de la clase en la evaluación y la proporción de estudiantes que acertó cada ítem, para que los docentes detecten los ítems especialmente difíciles para la clase.

Figura B.3: Plantilla del informe de aprendizaje, página 3: desempeño de la clase por tema y los estudiantes con necesidad de apoyo

Desempeño en cada tema:

Números y operaciones

- El estudiante típico de su clase respondió correctamente **4 de 5** preguntas en esta área.
- Dentro de esta área, sus estudiantes tienen más dificultades con la **suma de un número entero y un número decimal** y la **división de un número de dos dígitos por un número de un dígito**.
- Los tres estudiantes de su clase que más ayuda necesitan en esta área son:
 1. **Miton Das (N.º de lista 18, varón)**
 2. **Titas Das (N.º de lista 23, varón)**
 3. **Md Mehedi Hasan (N.º de lista 41, varón)**

Geometría y medición

- El estudiante típico de su clase respondió correctamente **1 de 3** preguntas en esta área.
- Dentro de esta área, sus estudiantes tienen más dificultades con la **suma de números y conversión de unidades de medida**.
- Los tres estudiantes de su clase que más ayuda necesitan en esta área son:
 1. **Mst Takiya Akter (N.º de lista 28, mujer)**
 2. **Israt Jahan Thrisha (N.º de lista 64, mujer)**
 3. **Md Abjal (N.º de lista 29, varón)**

Datos

- El estudiante típico de su clase respondió correctamente **1 de 2** preguntas en esta área.
- Dentro de esta área, sus estudiantes tienen más dificultades con la **interpretación de tablas con datos para calcular una fracción**.
- Los tres estudiantes de su clase que más ayuda necesitan en esta área son:
 1. **Mst Takiya Akter (N.º de lista 28, mujer)**
 2. **Mst Maisha Akter Maya (N.º de lista 7, mujer)**
 3. **Mst Tachlima Khandaker (N.º de lista 66, mujer)**

Notas: La figura muestra la tercera página del informe de aprendizaje. Desagrega el desempeño de la clase por tema (número, medición y geometría, y datos) y nombra a los estudiantes con mayor necesidad de apoyo en cada tema.

Figura B.4: Plantilla del informe de aprendizaje, página 4: resumen de la mejora de la clase y operaciones con necesidad de refuerzo

Mejoras desde junio de 2022:

¿Cómo mejoraron sus estudiantes?

- El estudiante típico de su clase respondió correctamente 1 pregunta más de las 10.
- En su clase, el **60%** respondió correctamente menos preguntas, el **16%** respondió correctamente más preguntas y el **24%** respondió correctamente la misma cantidad de preguntas.
- Los estudiantes de bajo desempeño de su clase respondieron correctamente 1 pregunta más de las 10.
- El área en la que los estudiantes de su clase mejoraron menos fue **medición y geometría**.
- Los temas en los que los estudiantes de su clase mejoraron menos fueron **aplicación de propiedades geométricas para hallar un ángulo faltante** y **suma de números y conversión de unidades de medida**.

Mejora en cada tema:

Números y operaciones

- El estudiante típico de su clase respondió correctamente 1 pregunta más de las 5 de esta área.
- Dentro de esta área, sus estudiantes tienen más dificultades con la **división de un número de dos dígitos por un número de un dígito** y la **división de un número de dos dígitos por un número de un dígito**.

Geometría y medición

- El estudiante típico de su clase respondió correctamente la misma cantidad de preguntas de las 3 de esta área.
- Dentro de esta área, sus estudiantes tienen más dificultades con la **aplicación de propiedades geométricas para hallar un ángulo faltante**.

Datos

- El estudiante típico de su clase respondió correctamente la misma cantidad de preguntas de las 2 de esta área.
- Dentro de esta área, sus estudiantes tienen más dificultades con la **interpretación de un calendario para hacer extrapolaciones**.

Notas: La figura muestra la cuarta página del informe de aprendizaje. Resume cómo cambió el desempeño de la clase desde la evaluación anterior (cuando corresponde) y señala las operaciones y procedimientos matemáticos que parecían necesitar refuerzo.

Figura B.5: Plantilla del informe de aprendizaje, página 5: usos posibles del informe

¿Qué debería hacer a continuación?

Esto es lo que debería hacer a continuación:

1. Use las listas de la página anterior para identificar a los estudiantes que más ayuda necesitan;
2. Use las copias adicionales de la prueba administrada en mayo de 2022 y pida a estos estudiantes que vuelvan a intentarla;
3. Revise las respuestas de cada estudiante e identifique las preguntas que respondió incorrectamente;
4. Pida a cada estudiante que le explique qué entendió que pedía la pregunta, cómo abordó la búsqueda de una solución y si se le ocurre otra forma de resolverla.
5. Use el documento de análisis de errores para explicar a los estudiantes por qué la opción que eligieron es incorrecta y por qué otra opción es la correcta.

Estos son algunos consejos para hacerlo:

- Los estudiantes pueden sentir vergüenza por sus errores, así que tranquilícelos: dígales que les pide rehacer la prueba para **entender en qué se equivocaron y cómo pueden mejorar**.
- El objetivo de este proceso es entender lo que el estudiante no sabe, así que **resista la tentación de ayudarlos** cuando vuelvan a hacer la prueba.
- Si un estudiante se traba, **pregúntele cuáles son las opciones** que está considerando. Si permanece en silencio, **ofrézcale una pista**, pero pregúntele si sabe por qué esa pista podría ser correcta.

Notas: La figura muestra la quinta y última página del informe de aprendizaje. Ofrece sugerencias concretas sobre cómo los docentes pueden usar la información del informe—por ejemplo, para agrupar a los estudiantes según sus necesidades, para repasar operaciones específicas en las clases siguientes o para comunicarse con las familias sobre el progreso.

Figura B.6: Guía de análisis de errores, ejemplo 1: suma de números decimales

1. ¿Cuál es la suma de 11.2 y 25?

- a. 3.62
- b. 13.6
- c. 13.8
- d. 36.2

La respuesta correcta es D: 36.2

- **Primero, el estudiante debe sumar los números en la posición de las unidades (es decir, 1 de 11 y 5 de 25) y anotar el resultado (es decir, 6).**
- **Luego, debe sumar los números en la posición de las decenas (es decir, 1 de 11 y 2 de 25) y anotar el resultado (es decir, 3).**
- **Luego, debe sumar los números de los decimales (es decir, 2 de 11.2 y 0 de 25.0) y anotar el resultado (es decir, 2).**
- **Con un 3 en la posición de las decenas, un 6 en la posición de las unidades y 2 en el decimal, la respuesta es 36.2**

Si un estudiante eligió...

- a. *Podría haber realizado el cálculo correctamente, pero haber cometido un error al entender cómo deben considerarse los decimales en las operaciones de suma.*
- b. *Podría haber reconocido mal la operación y sumado 11 y 2 del primer número, cometiendo además un error en los decimales.*
- c. *Podría haber realizado el cálculo correctamente, salvo por alinear mal los números sin considerar los dígitos de los decimales.*

Notas: La figura muestra el primer ejemplo de la guía de análisis de errores entregada a los docentes de tratamiento junto con los resultados de las evaluaciones. Para cada ítem, la guía enumera la respuesta correcta, el procedimiento necesario para alcanzarla y las concepciones erróneas más probablemente subyacentes a cada respuesta incorrecta. Se brindaron explicaciones similares para los 10 ítems de cada evaluación.

Figura B.7: Guía de análisis de errores, ejemplo 2: problema con enunciado de resta

3. Marie tiene 9 lápices. Le da 4 lápices a su hermana. ¿Cuántos lápices tiene Marie ahora?
- a. 9
 - b. 13
 - c. 4
 - d. 5

La respuesta correcta es D: 5.

- **Primero, el estudiante debe reconocer que se trata de una pregunta de resta en la que se resta 4 de 9. Debe prestar atención al orden de la resta y no confundirse.**
- **Luego, el estudiante debe restar 4 de 9 y anotar la respuesta 5.**
- **El estudiante debe escribir 5 como la respuesta correcta.**

Si un estudiante eligió...

- a. Podría haber copiado simplemente el número 9 directamente de la pregunta, sin reconocer ninguna operación.
- b. Podría haber interpretado que la operación era una suma en lugar de una resta.
- c. Podría haber copiado simplemente el número 4 directamente de la pregunta, sin reconocer ninguna operación.

Notas: La figura muestra el segundo ejemplo de la guía entregada a los docentes de tratamiento junto con los resultados de las evaluaciones. Ilustra cómo la guía abordaba problemas con enunciado contextualizados además de ítems procedimentales.

Tabla B.1: Marco de la evaluación de matemática de la intervención, por tema

(1) Tema	(2) Objetivo	(3) Real	(4) Temas evaluados	(5) Ítems
A. Número (objetivo: 50 %; real: 50 %)				
Números naturales	25 %	30 %	● Multiplicar y dividir números naturales, incluido el cálculo en problemas contextualizados simples. ● Resolver problemas que involucran números pares e impares, múltiplos y factores de números, redondeo de números y estimaciones.	2, 4 3
Expresiones, ecuaciones simples y relaciones	15 %	10 %	● Identificar o escribir expresiones o enunciados numéricos para representar situaciones problemáticas que pueden involucrar incógnitas.	5
Fracciones y decimales	10 %	10 %	● Demostrar conocimiento del valor posicional decimal, incluida la representación de decimales con palabras, números o modelos; comparar, ordenar y redondear decimales; sumar y restar decimales.	1
B. Medición y geometría (objetivo: 30 %; real: 30 %)				
Medición	15 %	10 %	● Resolver problemas que involucran masa, volumen y tiempo; identificar tipos y tamaños de unidades apropiados y leer escalas.	7
Geometría	15 %	20 %	● Identificar y dibujar rectas paralelas y perpendiculares; identificar y dibujar ángulos rectos y ángulos menores o mayores que un ángulo recto; comparar ángulos según su tamaño.	8
			● Usar propiedades elementales, incluidas la simetría axial y rotacional, para describir, comparar y crear figuras bidimensionales comunes.	9
C. Datos (objetivo: 20 %; real: 20 %)				
Lectura, interpretación y representación de datos	15 %	10 %	● Leer e interpretar datos de tablas, pictogramas, gráficos de barras, gráficos de líneas y gráficos circulares.	10
Uso de datos para resolver problemas	5 %	10 %	● Usar datos para responder preguntas que van más allá de la lectura directa de las presentaciones de datos, incluidos resolver problemas, combinar datos de múltiples fuentes y extraer conclusiones.	6

Notas: La tabla muestra la distribución objetivo y la real de los ítems por tema en las evaluaciones de matemática de la intervención de 10 ítems administradas en mayo y agosto. El marco se basó en el currículo de 1.º a 5.º grado de Bangladesh y en un marco de evaluación internacional de matemática de nivel primario. Los ítems se referencian por su posición en la evaluación.

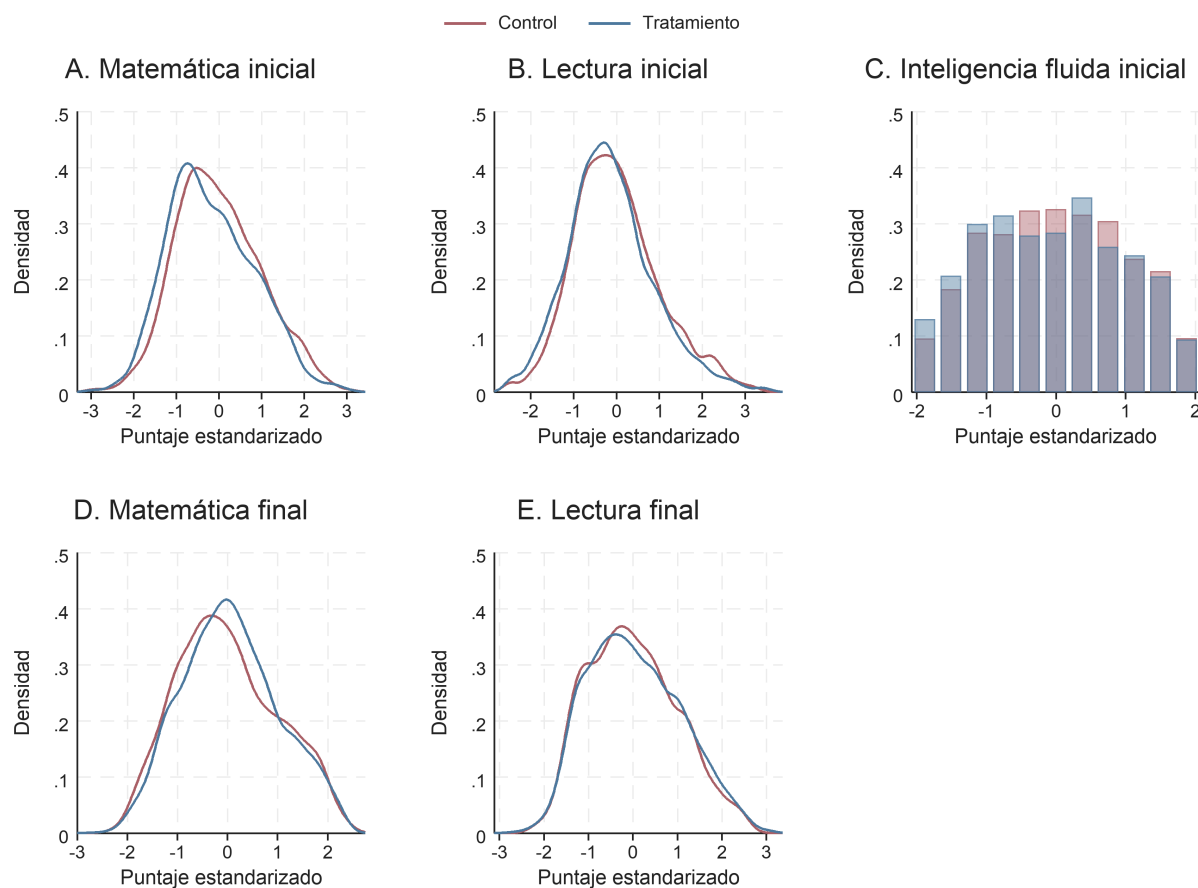
Tabla B.2: Marco de la evaluación de matemática de la intervención, por dominio cognitivo

(1) Dominio cognitivo	(2) Objetivo	(3) Real	(4) Habilidades evaluadas	(5) Ítems
<i>A. Conocer (objetivo: 40 %; real: 40 %)</i>				
Recordar	10 %	10 %	• Recordar definiciones, terminología, propiedades de los números, unidades de medida, propiedades geométricas y notación.	9
Calcular	30 %	30 %	• Ejecutar procedimientos algorítmicos de suma, resta, multiplicación, división o una combinación de estos con números naturales, fracciones, decimales y enteros; ejecutar procedimientos algebraicos sencillos.	1, 2, 4
<i>B. Aplicar (objetivo: 40 %; real: 40 %)</i>				
Determinar	10 %	10 %	• Determinar operaciones, estrategias y herramientas eficientes o apropiadas para resolver problemas para los que existen métodos de solución de uso corriente.	8
Implementar	30 %	30 %	• Implementar estrategias y operaciones para resolver problemas que involucran conceptos y procedimientos matemáticos conocidos.	3, 7, 10
<i>C. Razonar (objetivo: 20 %; real: 20 %)</i>				
Analizar	10 %	10 %	• Determinar, describir o usar relaciones entre números, expresiones, cantidades y figuras.	5
Integrar/Sintetizar	10 %	10 %	• Vincular distintos elementos del conocimiento, representaciones relacionadas y procedimientos para resolver problemas.	6

Notas: La tabla muestra la distribución objetivo y la real de los ítems por habilidad cognitiva en las evaluaciones de matemática de la intervención de 10 ítems. Los objetivos de subhabilidad se calculan a partir de la proporción de ítems de cada subhabilidad (el marco especifica objetivos solo a nivel de habilidad—Conocer, Aplicar y Razonar—por lo que fijamos los objetivos de subhabilidad iguales a su proporción de ítems efectiva). Los ítems se referencian por su posición en la evaluación.

Apéndice C Medición

Figura C.1: Distribuciones de las evaluaciones de estudiantes (línea de base y línea de cierre)



Notas: La figura muestra gráficos de densidad kernel de los puntajes de matemática y lectura e histogramas de los puntajes de inteligencia fluida, por separado según el grupo experimental. Los paneles A–C muestran los puntajes de línea de base; los paneles D y E muestran los puntajes de cierre. Matemática y lectura están escaladas por TRI y estandarizadas respecto de la distribución del grupo de control dentro de cada ronda; los puntajes de inteligencia fluida están estandarizados respecto de la distribución del grupo de control de línea de base.

Tabla C.1: Marco de la evaluación de impacto de matemática, por tema

	(1)	(2)	(3)	(4)	(5)
Tema		Objetivo	Real	Cantidad de ítems	Ítems
<i>A. Número (objetivo: 50 %; real: 47 %)</i>					
Números naturales		25 %	27 %	8	1-8
Valor posicional		0 %	0 %	0	—
Sumar y restar		10 %	10 %	3	1, 3, 4
Multiplicar y dividir		7 %	7 %	2	2, 6
Números pares/impares, factores, redondeo y estimaciones		3 %	3 %	1	5
Propiedades de números y operaciones		7 %	7 %	2	7, 8
Expresiones, ecuaciones simples y relaciones		15 %	13 %	4	9-12
Número u operación faltante		3 %	3 %	1	9
Expresiones o enunciados numéricos		3 %	3 %	1	10
Patrones y relaciones		7 %	7 %	2	11, 12
Fracciones y decimales		10 %	7 %	2	13, 14
Fracciones		3 %	3 %	1	13
Decimales		3 %	3 %	1	14
<i>B. Medición y geometría (objetivo: 30 %; real: 30 %)</i>					
Medición		15 %	17 %	5	15-19
Longitudes		7 %	7 %	2	15, 19
Masa, volumen, tiempo, unidades y escalas		7 %	7 %	2	16, 17
Perímetro, área y volumen		3 %	3 %	1	18
Geometría		15 %	13 %	4	20-23
Rectas y ángulos		3 %	3 %	1	20
Propiedades de las figuras y simetría		3 %	3 %	1	21
Patrones y relaciones		7 %	7 %	2	22, 23
<i>C. Datos (objetivo: 20 %; real: 23 %)</i>					
Lectura, interpretación y representación de datos		15 %	17 %	5	24-28
Leer e interpretar presentaciones de datos		7 %	7 %	2	24, 25
Organizar y representar datos		10 %	10 %	3	26-28
Uso de datos para resolver problemas		5 %	7 %	2	29, 30
Usar datos más allá de la lectura directa		7 %	7 %	2	29, 30

Notas: La tabla muestra la distribución objetivo y la real de los ítems en las evaluaciones de impacto de matemática de línea de base y de cierre, por tema. Las proporciones objetivo provienen del marco de evaluación cuando están disponibles; en caso contrario, las proporciones objetivo de los subdominios se fijan iguales a las proporciones de ítems efectivas porque el marco no especifica objetivos separados para esos subdominios. Las proporciones reales y las posiciones de los ítems se calculan a partir de las listas de ítems de los libros de trabajo del marco de línea de base y de cierre, que usan la misma asignación. Los ítems se referencian por su posición en cada evaluación.

Tabla C.2: Marco de la evaluación de impacto de matemática, por dominio cognitivo

(1)	(2)	(3)	(4)	(5)
Dominio cognitivo	Objetivo	Real	Cantidad de ítems	Ítems
<i>A. Conocer (objetivo: 40 %; real: 43 %)</i>				
Recordar	13 %	13 %	4	3, 12, 21, 24
Reconocer	7 %	7 %	2	6, 20
Clasificar/ordenar	0 %	0 %	0	—
Calcular	13 %	13 %	4	1, 2, 14, 29
Recuperar	3 %	3 %	1	26
Medir	7 %	7 %	2	17, 18
<i>B. Aplicar (objetivo: 40 %; real: 37 %)</i>				
Determinar	20 %	20 %	6	5, 7, 9, 13, 27, 30
Representar/modelar	7 %	7 %	2	22, 23
Implementar	10 %	10 %	3	4, 15, 16
<i>C. Razonar (objetivo: 20 %; real: 20 %)</i>				
Analizar	10 %	10 %	3	10, 25, 28
Integrar/sintetizar	3 %	3 %	1	11
Evaluar	7 %	7 %	2	8, 19
Extraer conclusiones	0 %	0 %	0	—
Generalizar	0 %	0 %	0	—
Justificar	0 %	0 %	0	—

Notas: La tabla muestra la distribución objetivo y la real de los ítems en las evaluaciones de impacto de matemática de línea de base y de cierre, por dominio cognitivo. Las proporciones objetivo de nivel superior provienen del marco de evaluación; las proporciones objetivo de subhabilidad se fijan iguales a las proporciones de ítems efectivas porque el marco no especifica objetivos separados para subhabilidades individuales. Las proporciones reales y las posiciones de los ítems se calculan a partir de las listas de ítems de los libros de trabajo del marco de línea de base y de cierre, que usan la misma asignación. Los ítems se referencian por su posición en cada evaluación.

Tabla C.3: Marco de la evaluación de lectura, por habilidad lectora

(1)	(2)	(3)	(4)	(5)
Dominio/área	Objetivo	Real	Habilidades evaluadas	Ítems
<i>A. Vocabulario (objetivo: 50 %; real: 50 %)</i>				
Vinculación de imágenes		20 %	Identificar imágenes que se corresponden con palabras dadas u oraciones cortas; mide habilidades de lenguaje receptivo.	1-6
Tarea de completamiento		3 %	Oración con una palabra parcial o totalmente cubierta; los estudiantes usan el contexto de la oración para identificar la palabra faltante.	7
Tarea de elección múltiple en contexto		27 %	Oración con dos o tres palabras alternativas; los estudiantes seleccionan la palabra que tiene más sentido en la oración.	12-15, 19-22
<i>B. Comprensión (objetivo: 50 %; real: 50 %)</i>				
Localizar y recuperar		10 %	Localizar y recuperar información explícitamente enunciada en el texto.	16-18
Realizar inferencias		13 %	Realizar inferencias directas a partir de la información del texto.	8-11
Interpretar e integrar		10 %	Interpretar e integrar ideas e información a lo largo del texto.	28-30
Evaluar y criticar		17 %	Evaluar y criticar el contenido y los elementos textuales.	23-27

Notas: La tabla muestra la distribución objetivo y la real de los ítems por habilidad lectora en las evaluaciones de lectura de 30 ítems de línea de base y de cierre. Los objetivos de nivel superior (Vocabulario y Comprensión) se especifican a nivel de dominio; el marco no especifica objetivos separados para subhabilidades individuales, por lo que la columna Objetivo de las subhabilidades se deja en blanco. Los ítems se referencian por su posición en la evaluación.

Tabla C.4: Marco de la evaluación de lectura, por nivel de insumo

(1)	(2)	(3)	(4)	(5)
Nivel de insumo	Objetivo	Real	Descripción	Ítems
Palabras	20 %	20 %	Los ítems presentan una única palabra o frase corta a la que los estudiantes deben responder (por ejemplo, vincular una palabra con una imagen).	1-6
Oraciones	20 %	20 %	Los ítems presentan una o pocas oraciones a las que los estudiantes deben responder (por ejemplo, completar una oración, vincular una oración con una imagen).	7, 16-18, 23-24
Párrafos	40 %	40 %	Los ítems siguen a un párrafo corto (típicamente de varias oraciones) sobre el cual los estudiantes responden preguntas de comprensión o de completamiento.	8-15, 19-22
Relato	20 %	20 %	Los ítems siguen a un texto narrativo o informativo más extenso que abarca múltiples párrafos.	25-30

Notas: La tabla muestra la distribución objetivo y la real de los ítems por nivel de insumo en las evaluaciones de lectura de 30 ítems. El nivel de insumo se refiere a la extensión y la estructura del texto en el que se basa el ítem, que va desde una única palabra hasta un relato de varios párrafos. Los ítems se referencian por su posición en la evaluación.

Tabla C.5: Componentes de varianza de los puntajes de interacción entre estudiantes y docentes, muestra completa (línea de cierre)

	(1) Iniciador	(2) Dirección	(3) Contenido	(4) Tono	(5) Formulación	(6) Duración
Componente de varianza	Var.	Var.	Var.	Var.	Var.	Var.
Docente	0.05	0.10	0.03	0.15	0.07	0.25
Residual	0.16	0.20	0.12	0.17	0.15	0.32
DE del efecto del docente	0.22	0.32	0.16	0.39	0.26	0.50
EEM de una única observación	0.40	0.44	0.34	0.41	0.39	0.57
Confiabilidad de una única observación						
Posición relativa de los docentes	0.24	0.35	0.18	0.48	0.30	0.43
Puntajes absolutos de los docentes	0.24	0.35	0.18	0.48	0.30	0.43
N (docentes)	156	156	156	156	156	156

Notas: La tabla muestra los componentes de varianza de un estudio de G de los puntajes de interacción entre estudiantes y docentes en la muestra completa. El objeto de medición es el docente y la faceta de error es la ocasión, definida como una de las primeras diez interacciones puntuadas entre estudiantes y docentes de la clase. El DE del efecto del docente es la raíz cuadrada de la varianza del docente. El error estándar de medición (EEM) de una única observación es la raíz cuadrada de la varianza de error relativa. Los componentes de varianza estimados como negativos se fijan en cero.

Tabla C.6: Componentes de varianza de los puntajes de interacción entre estudiantes y docentes, submuestra de dos evaluadores (línea de cierre)

	(1)	(2)	(3)	(4)	(5)	(6)
	Iniciador	Dirección	Contenido	Tono	Formulación	Duración
Componente de varianza	Var.	Var.	Var.	Var.	Var.	Var.
Docente	0.04	0.04	0.01	0.08	0.07	0.22
Evaluador	0.00	0.00	0.00	0.00	0.00	0.00
Ocasión $\times$ docente	0.15	0.16	0.13	0.08	0.11	0.28
Evaluador $\times$ docente	0.00	0.01	0.01	0.00	0.00	0.00
Residual	0.04	0.05	0.02	0.07	0.03	0.06
DE del efecto del docente	0.21	0.19	0.11	0.28	0.27	0.47
EEM de una única observación	0.13	0.16	0.11	0.20	0.13	0.18
Confiabilidad de una única observación						
Posición relativa de los docentes	0.70	0.58	0.52	0.67	0.81	0.87
Puntajes absolutos de los docentes	0.20	0.16	0.08	0.41	0.36	0.41
N (docentes)	48	48	48	48	48	48

Notas: La tabla muestra componentes de varianza de un estudio de G de los puntajes de interacción entre estudiantes y docentes en la submuestra aleatoria de dos evaluadores. El objeto de medición es el docente y las facetas de error son el evaluador y la ocasión, definida como una de las primeras diez interacciones puntuadas entre estudiantes y docentes de la clase. El DE del efecto del docente es la raíz cuadrada de la varianza del docente. El error estándar de medición (EEM) de una única observación es la raíz cuadrada de la varianza de error relativa. Los componentes de varianza estimados como negativos se fijan en cero.