

Complementing Teachers: Can Artificial Intelligence Improve Student Learning by Addressing Learning Variability?

Pre-Analysis Plan

Felipe Barrera-Osorio and Juan Muñoz-Morales *

December 10, 2025

Abstract

Education systems in developing countries face a persistent learning crisis characterized by low average achievement and substantial disparities in student learning. This paper evaluates a teacher-level, artificial intelligence-enabled intervention designed to support lesson planning and promote within-classroom differentiated instruction. We implement a school-level randomized controlled trial in low-income public secondary schools in Bogotá, Colombia. The intervention provides mathematics teachers with access to an AI-based lesson-planning platform that integrates learner variability and growth-oriented pedagogical frameworks, reducing preparation costs while improving instructional alignment with heterogeneous student needs. We assess impacts on student learning, with particular attention to the heterogeneity of the effect in a context marked by pronounced educational inequality. Using baseline and endline student assessments and complementary teacher surveys, we estimate both intention-to-treat and treatment-on-the-treated effects, and examine treatment effect heterogeneity using pre-specified subgroup analyses and data-driven methods.

Keywords: Artificial Intelligence, Classrooms, Math, Treatment Effect Heterogeneity, Colombia.

JEL classifications: C93, I21, J16, J24, O15, O33.

Study pre-registration: [AEARCTR-0017147](https://www.aearctr.org/0017147)

*Barrera-Osorio: Vanderbilt University (felipe.barrera.-osorio@vanderbilt.edu). Muñoz-Morales: IESEG School of Management, Univ. Lille, CNRS, UMR 9221- LEM-Lille Économie Management, F-59000 Lille, France (j.munoz@ieseg.fr). We acknowledge the financial support of the partnership for Tech in Education (P4T-Ed) initiative by Innovations for Poverty Action (IPA) and the Jacobs Foundation.

1 Introduction

Education systems in developing countries face a widely documented “learning crisis”, characterized by low levels of learning despite substantial gains in school enrollment (World Bank, 2018). Learning outcomes remain particularly weak among disadvantaged students, generating large and persistent achievement gaps that reinforce economic inequality (Bassi et al., 2015). For example, students in Latin America are estimated to require more than two centuries to reach the average reading proficiency observed among students in OECD countries. These disparities have been further exacerbated by the COVID-19 pandemic, which disproportionately reduced learning among poorer students in low- and middle-income countries. Addressing this crisis has therefore become a central policy objective, motivating a growing body of interventions aimed at improving instructional quality and student learning.

A prominent strand of this literature evaluates the use of educational technology to enhance teacher effectiveness and student achievement. While evidence on the impacts of such interventions is mixed, a common feature is their assignment at the classroom, school, or teacher level, implicitly treating classrooms as homogeneous units. Yet student learning varies substantially within classrooms, reflecting differences in prior knowledge, socioeconomic background, and learning pace. As a result, interventions that raise average classroom performance may leave the lowest-performing students behind. An alternative approach assigns interventions at the student level, tailoring instruction to individual needs. However, many such programs operate outside regular school hours, raising concerns about scalability, sustainability, and the efficient use of children’s time. These limitations underscore the need for interventions that address within-classroom heterogeneity while operating through teachers during standard instructional time.

Recent advances in artificial intelligence (AI) offer a novel opportunity to bridge this gap by enabling teachers to better diagnose and respond to variation in student learning within their classrooms. AI-based tools can support differentiated instruction by providing real-time feedback, adaptive content, and targeted pedagogical guidance, potentially enhancing teachers’ capacity to serve students at different achievement levels. Despite this promise, rigorous evidence on the effectiveness of AI-enabled, teacher-level interventions—particularly for improving outcomes among lower-achieving and disadvantaged students—remains scarce.

This paper evaluates the impact of a artificial intelligence-enabled, teacher-level intervention designed to address within-classroom heterogeneity in student learning. We partner with *Mentu*, a Colombian education technology startup specializing in AI-based learning tools for low-income students, and with the *Secretaría de Educación Distrital* (SED), the public education authority in Bogotá, Colombia. Together, we de-

signed and implemented an intervention that provides participating schools with access to ShaIA, an AI-driven platform that supports teachers by delivering targeted, evidence-based guidance for lesson planning and classroom instruction.

ShaIA is built around the integration of two complementary pedagogical frameworks: the learner variability model (Pape, 2018) and the math mindsets approach (Boaler and Dweck, 2016).¹ By combining these frameworks, the platform generates structured recommendations that encourage differentiated instructional strategies tailored to heterogeneous student needs within the same classroom. The intervention operates by reducing the fixed costs of high-quality lesson preparation—a key constraint faced by teachers in under-resourced settings. By automating and organizing the planning process, ShaIA allows teachers to devote less time to preparation while increasing the pedagogical relevance and adaptability of their instructional choices. In doing so, the platform aims to enhance teacher engagement and instructional effectiveness, particularly among low-value-added teachers who often lack access to ongoing feedback and pedagogical support. These features make the intervention especially relevant in public schools serving low-income communities in Bogotá, where within-classroom learning gaps are large and institutional support for differentiated instruction is limited.

Importantly, the intervention is implemented in a context characterized by substantial learning disparities across student demographic groups, particularly by gender and migration status. Colombia exhibits one of the largest gender gaps in mathematics achievement globally (Muñoz, 2018). In addition, the country has experienced large inflows of displaced populations over the past decades, including approximately nine million internally displaced individuals due to armed conflict and nearly three million Venezuelan migrants in recent years. By explicitly incorporating learner variability into instructional planning, the intervention is expected to generate larger learning gains among historically disadvantaged groups—particularly female and migrant students—whose needs are often underserved by uniform, classroom-average teaching practices. We complement this analysis by formally testing the null hypothesis of a common treatment effect (Buhl-Wiggers et al., 2024) and estimating conditional average treatment effects following Wager and Athey (2018). This data-driven analysis will help identify the subgroups that benefit the most from the intervention and contrast them with our pre-specified hypothesis about treatment heterogeneity.

The evaluation is originally conducted in 158 eligible public schools located in low-income areas of Bogotá, which predominantly serve disadvantaged students with

¹The learner variability framework, developed by Digital Promise Global, emphasizes that students differ systematically across cognitive, social, and contextual dimensions, and that instruction should be designed to accommodate this variation. The math mindsets approach draws on recent evidence from the science of learning to promote effective mathematics instruction through growth-oriented pedagogical practices.

limited access to high-quality educational resources. Mathematics teachers in 58 secondary schools (grades 6 through 9) were offered access to ShaIA, together with complementary workshops designed to promote effective use of the platform. The intervention was rolled out through a school-level randomized design, in which all mathematics teachers in treated schools were offered access to the program.

Participation was voluntary for both schools and teachers. Therefore, to implement the randomization, we employed a block randomization strategy, grouping schools into 52 blocks based on pre-intervention characteristics. Within each block, access to the platform was randomly assigned, ensuring comparability between treatment and control schools. While some schools assigned to treatment chose not to participate, non-participation within a given block does not compromise the randomization in other blocks. Random assignment was preserved in 26 blocks, resulting in a balanced sample of 40 treated and 50 control schools. Given voluntary take-up among teachers in treated schools, our empirical strategy estimates both intention-to-treat and treatment-on-the-treated effects.

Student learning outcomes were measured at baseline in August 2025, with endline assessments scheduled for the end of the academic year, prior to December 2026. In addition, we administered teacher surveys in both treatment and control schools to examine changes in instructional practices, pedagogical beliefs, and engagement.

This paper contributes to several strands of the economics of education literature. First, it adds to the growing body of work studying the effects of technology in education, and more specifically to recent research examining artificial intelligence as an input to the production of human capital ([Oreopoulos et al., 2024](#); [De Simone et al., 2025](#); [Lopez et al., 2025](#); [Andrabi et al., 2025](#); [Bastani et al., 2024](#); [Henkel et al., 2024](#); [Lehmann et al., 2024](#)). While this literature documents mixed effects of educational technology (and also AI) on student learning ([Rodriguez-Segura, 2021](#); [Escueta et al., 2017](#)), relatively little is known about whether AI-based tools can improve learning by enhancing teachers' ability to address within-classroom heterogeneity.²

Second, this paper speaks to the distinction between technology-based interventions implemented at the teacher or classroom level and those implemented at the student level. Teacher- or classroom-level interventions typically target average classroom outcomes and abstract from within-classroom variation, yielding mixed evidence of effectiveness ([Jackson and Makarin, 2018](#); [Barrow et al., 2009](#); [Angrist and Lavy, 2002](#); [Berlinski and Busso, 2017](#)). Student-level interventions, by contrast, often find positive effects when providing individualized instruction or adaptive learning during supplemental instructional time ([Banerjee et al., 2007](#); [Barrow et al., 2009](#); [Mu-](#)

²[Rodriguez-Segura \(2021\)](#) and [Escueta et al. \(2017\)](#) provide comprehensive reviews of the mixed evidence on technology-driven education interventions in developing and developed countries, respectively.

ralidharan et al., 2019; Ma et al., 2024; Büchel et al., 2022; Blimpo et al., 2020), but negative or null effects when technology substitutes for classroom instruction, such as through one-to-one laptop programs (Beuermann et al., 2015; Cristia et al., 2017; Berlinski and Busso, 2017). Unlike either approach, our intervention leverages recent advances in AI to promote differentiated instruction within the classroom during regular school hours, operating through teachers rather than bypassing them. By doing so, the paper provides novel evidence on whether AI-enabled pedagogical support can reduce learning gaps and improve outcomes for disadvantaged students by addressing a key mechanism underlying persistent educational inequality.

Third, this paper contributes to the literature on *teaching at the right level*, which emphasizes the importance of aligning instruction with students' actual learning levels. Existing evidence demonstrates substantial gains from interventions that tailor instruction through out-of-class tutoring (Banerjee et al., 2007, 2017), the use of volunteer instructors (Banerjee et al., 2010), and classroom tracking (Duflo et al., 2011). While effective, these approaches typically rely on additional instructional time, external personnel, or structural changes to classroom organization, raising concerns about scalability and integration into standard schooling systems. In contrast, none of these interventions explicitly focus on strengthening teachers' capacity to address learner variability within regular classroom instruction. By equipping teachers with AI-enabled tools to adapt pedagogy to heterogeneous student needs during school hours, our intervention offers a potentially more scalable and institutionally embedded approach to teaching at the right level (De Simone et al., 2025).

Finally, this paper contributes to the literature on heterogeneous treatment effects in education. While randomized evaluations have traditionally focused on average treatment effects, recent work shows that treatment effect heterogeneity is often substantial and comparable in magnitude to mean impacts (Buhl-Wiggers et al., 2024; Han et al., 2026). Understanding this heterogeneity is crucial for identifying mechanisms and improving policy targeting. Such variation can be studied through pre-specified subgroup analyses motivated by theory or policy relevance, as well as through data-driven methods that flexibly uncover heterogeneity across individuals or groups (Wager and Athey, 2018; Chernozhukov et al., 2022). We contribute to this literature by examining how an intervention explicitly designed to address within-classroom learning variation affects outcomes across pre-specified subgroups—particularly by gender and migration status—and by contrasting these results with estimates obtained from data-driven approaches.

2 Conceptual Framework

2.1 Learning Variability under Student Sorting

Consider a continuum of students indexed by i who are endowed with innate ability $\theta_i \sim N(\mu_\theta, \sigma_\theta^2)$. As students age, they sort into classrooms whose teachers vary in quality and is captured by $\gamma_c \sim N(\mu_\gamma, \sigma_\gamma^2)$. Students with different ability are allowed to sort into classrooms of different quality, implying that $\text{Cov}(\theta_i, \gamma_c) = \sigma_{\theta, \gamma}$.³ Hence, student ability and classroom/teacher quality are jointly normally distributed:

$$\begin{pmatrix} \theta \\ \gamma \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\theta \\ \mu_\gamma \end{pmatrix}, \begin{pmatrix} \sigma_\theta^2 & \sigma_{\theta, \gamma} \\ \sigma_{\theta, \gamma} & \sigma_\gamma^2 \end{pmatrix}\right).$$

Student learning is determined by both ability and teacher quality:

$$y_{ic} = \theta_i + \gamma_c,$$

which implies that $\mathbb{E}[y] = \mu_\theta + \mu_\gamma$, and $\text{Var}(y) = \sigma_y^2 = \sigma_\theta^2 + \sigma_\gamma^2 + 2\sigma_{\theta, \gamma}$.

Under sorting, a student i who enters a classroom whose teachers is of quality γ_c is expected to learn:

$$\mathbb{E}[y \mid \gamma_c] = (\mu_\theta + \mu_\gamma) + \left(\frac{\sigma_\gamma^2 + \sigma_{\theta, \gamma}}{\sigma_\gamma^2}\right) (\gamma_c - \mu_\gamma).$$

Learning therefore depends on student ability, teacher quality, and the degree of assortative sorting. If students do not selectively sort across classrooms (i.e., $\sigma_{\theta, \gamma} = 0$), then $\mathbb{E}[y] = \mathbb{E}[y \mid \gamma_c]$. When selection is positive ($\sigma_{\theta, \gamma} > 0$) differences in innate ability are amplified by sorting.

Learner variability also depends on the extent of sorting. We can decompose total learning variation, $\text{Var}(y) = \sigma_y^2$, into *within* and *between* classroom components:

$$\begin{aligned} \text{Var}(y) &= \mathbb{E}[\text{Var}(y \mid \gamma)] + \text{Var}(\mathbb{E}[y \mid \gamma]) \\ &= \underbrace{\frac{\sigma_\gamma^2 \sigma_\theta^2 - \sigma_{\theta, \gamma}^2}{\sigma_\gamma^2}}_{\text{Within}} + \underbrace{\frac{(\sigma_\gamma^2 + \sigma_{\theta, \gamma})^2}{\sigma_\gamma^2}}_{\text{Between}}. \end{aligned}$$

Between-classroom variation depends positively on student sorting. It often arises

³For simplicity, we assume that each classroom is taught by a single teacher. Our framework further assumes that school-level sorting, a key source of heterogeneity, is fully captured by the assignment of students to classrooms. This entails the additional assumption that classrooms and teachers within the same school are homogeneous, such that all within-school variation in instruction is attributed to differences across classrooms rather than within them.

from social constraints (e.g., geographic or income-based) that induce students to select into different classrooms that vary in quality, typically generating a positive correlation between student ability and teacher quality. *Within*-classroom heterogeneity, in contrast, depends negatively on sorting: when positive sorting is strong, variation within classrooms decreases because students in the same classroom tend to be more similar.

2.2 Targeting Learning Variation

Reducing learning inequality requires addressing both *between* and *within* variation, since policies that reduce between-classroom variation may exacerbate within-classroom heterogeneity. In fact, when sorting decreases (e.g., under random assignment), between-classroom heterogeneity mechanically falls, but within-classroom heterogeneity rises. Thus, policies that limit sorting and promote fairer allocation may increase within-classroom variation, making it even more important to address it directly.

However, addressing within-classroom heterogeneity is particularly challenging for at least two reasons. First, students in a classroom share limited resources that must be allocated across them. For example, a teacher—usually assigned at the classroom level—must divide their time and attention across all students, making it difficult to individually target students with specific needs.

Second, within-classroom variation could in principle be reduced either by lowering learning among top-performing students or by improving learning among low-performing students. The former is undesirable, while the latter is the intended goal. However, targeted interventions for low performers are costly, and—because resources are limited and shared—may reduce the support available to higher-performing students. A trade-off thus naturally emerges.

Existing evidence has shown that teacher-level interventions are effective in improving student learning, as teachers constitute one of the most important inputs in the education production function (Rockoff, 2004; Rivkin et al., 2005; Hanushek, 2011; Chetty et al., 2011; Hanushek and Rivkin, 2012; Chetty et al., 2014; Araujo et al., 2016; Jackson, 2018; Rose et al., 2022). These policies directly affect classroom quality and therefore primarily reduce *between*-classroom variation. However, they do little to address the *within*-classroom heterogeneity highlighted above. As a result, even though teacher-level interventions can raise overall learning and narrow differences between classrooms, it remains unclear how their effectiveness can be translated into reducing within-classroom variation or supporting students who lag behind.

At the same time, interventions providing educational technologies have become increasingly common. The evidence on their effects, however, is mixed. On the one

hand, technologies that target students individually and operate independently of teachers have often produced negligible or even negative impacts on learning (Beuermann et al., 2015; Angrist and Lavy, 2002; Cristia et al., 2017; Berlinski and Busso, 2017; Ma et al., 2024). On the other hand, technologies that complement teachers' instruction have consistently raised learning outcomes across a variety of settings (Banerjee et al., 2007; Büchel et al., 2022). These complementary tools are most effective when they diversify instruction and enable students to learn at their own level (Banerjee et al., 2017; Muralidharan et al., 2019). Yet, many existing technologies fail to achieve this goal because teachers cannot easily use them to manage or respond to the heterogeneity within their classrooms.

The recent surge in AI-based tools creates new opportunities to overcome longstanding limitations in education. Advances in AI make it possible to design technologies that directly support teachers in addressing learning variability *within* the classroom. With broader access to these tools, interventions can provide educators with resources that help them manage student heterogeneity more effectively, delivering instruction tailored to each student's level. This approach enables students to progress within their individual zones of proximal development. Unlike earlier technologies, which often replaced or bypassed teachers, AI-based tools can be designed to complement teachers' work and enhance learning for all students.

These developments could also influence variation *between* classrooms. If lower-achieving classrooms benefit more quickly, gaps in achievement may narrow. Conversely, if higher-achieving classrooms gain disproportionately, differences could widen, increasing between-classroom variation. Despite the potential of AI to shape both *within-* and *between-*classroom learning, research on these effects remains limited.

2.3 Research Questions

Driven by this framework, this study seeks to answer the following questions.

1. Does the provision of AI-driven algorithms lead to changes in teachers' teaching practices?
2. Do AI-driven technologies increase teacher value-added translated into students' learning?
3. Do these technologies help address within classroom heterogeneity by improving traditionally low-performers?
4. Do the effects differ for students who are traditionally disadvantaged in math (such as migrants and women)?

5. Does the provision of AI-driven algorithms reduce both within and between classroom learning variation?

2.4 Hypothesis

Our central hypothesis is that AI can improve overall student learning while disproportionately benefiting groups that have historically lagged behind. Providing teachers with access to a well-designed AI ecosystem can enhance their teaching through two steps. First, AI can reduce the cost of class preparation and pedagogy delivery. In turn, AI tools that reduce lesson-planning time may increase teacher engagement and encourage consistent use of the platform. Second, by doing so, these tools promote more inclusive and adaptive teaching practices, enabling teachers to address differentiated learning needs without disadvantaging higher-performing students. By shifting teaching practices through these mechanisms, we expect teachers to more effectively manage within-classroom heterogeneity and better support low-performing learners.

Interventions designed to help teachers address within-classroom learning variability have the potential to close persistent achievement gaps that disproportionately affect certain student groups. Because AI can enhance teachers' capacity to tailor instruction to diverse learning profiles, we expect it to generate larger treatment effects for historically disadvantaged groups and thereby reduce within-classroom variability in learning outcomes. As discussed in Section 3.1, in the setting we analyze, girls and migrant students exhibit significantly lower baseline performance in math. We hypothesize that the intervention will allow teachers to better support these groups while raising overall learning. We will test this prediction by examining both average treatment effects and heterogeneity in impacts across student subgroups.

3 Setting

3.1 Background

Primary and secondary education in Colombia consists of three years of preschool, five years of primary school, four years of lower secondary, and two years of upper secondary. Education is provided by both public and private institutions. Approximately 80 percent of schools are public, and they enroll about 75 percent of the country's students. School attendance is compulsory until age 15, at which time students typically complete the lower secondary cycle. Students who finish the upper secondary cycle are required to take a national high school exit exam, known as Saber 11, in order to obtain their diploma. Those who graduate may then continue into tertiary education, which is also offered through a mix of public and private institutions.

Secondary school graduation rates have increased in recent decades (Bassi et al., 2015), but learning outcomes remain very low and highly unequal. Estimates indicate that roughly 70 percent of Colombian students do not reach the minimum level of math proficiency required to participate effectively in society, compared with an average of 22 percent in OECD countries (Icfes, 2024). Moreover, substantial disparities emerge across students from different backgrounds. Table 1 reports scores on the high school exit exam, disaggregated by type of school (public vs. private), gender, and migrant status.⁴

Low income children, women, and migrants score significantly lower, driving the overall learning rates to very low values. For instance, the gap between students in public and private in math is of about -0.47 standard deviations, implying significant income disparities since low income students are the most likely to attend public institutions. In addition, Colombia has been shown to be one of the countries with the largest math gender gap against females in mathematics (0.25 standard deviations) (Muñoz, 2018). A significant gap of -0.22 standard deviations in math also exists between local and migrant students, implying that the current surges in immigration have further decreased the learning averages in country.

Due to these persistent learning challenges, teachers have become a key policy focus for local governments, as they are central drivers of student achievement (Chetty et al., 2014; Hanushek and Rivkin, 2006; Rivkin et al., 2005). This growing interest led the national government to implement, in 2005, a large-scale merit-based reform that increased teacher salaries and introduced an entry exam for the public teaching profession. The aim of this sweeping reform was to attract a stronger pool of teachers, which would eventually translate into improved student learning outcomes.

However, implementation challenges prevented the reform from yielding positive effects on learning, raising serious concerns about the quality of public education in the country (Busso et al., 2024). The COVID-19 crisis further deepened these problems by disproportionately harming the learning of the most vulnerable students and widening existing gaps (Melo-Becerra et al., 2021). As a result, learning levels in Colombia remain low and highly unequal, and local governments continue to search for policies capable of reversing these trends.

3.2 Intervention

The low and unequal learning outcomes in the country have prompted local authorities to implement policies aimed at improving student achievement. In partner-

⁴Since 2015, Colombia has received approximately 8 million Venezuelan migrants, 40 percent of whom are under 18. A large share of these children have been successfully integrated into the education system and now account for roughly 10 percent of total school enrollment (Vargas and Rozo Villarraga, 2024).

ship with the *Secretaría de Educación Distrital* (SED)—the local education authority in Bogotá, Colombia—, and *Mentu*—a Colombian startup that provides AI-driven learning technologies—, we designed an intervention intended to enhance teachers’ value-added while addressing the heterogeneous learning needs of students from diverse backgrounds. The intervention targeted teachers working in low-income areas of the city and consisted of two components: (i) access to *ShaIA*, an AI-based pedagogical support platform designed to address learning variability; and (ii) a structured training program to support teachers in the effective use of the platform.

Access to AI-based pedagogical support. *ShaIA* is an AI-driven pedagogical ecosystem that provides personalized guidance to teachers. The platform is designed to support mathematics instruction by assisting teachers in planning and implementing lessons aligned with curricular standards and students’ diverse learning needs. Treated teachers were granted access to the platform, where they created a course profile—describing course characteristics to the algorithm—and specified learning objectives in mathematics.

ShaIA generates recommendations by channeling a large language model guided by two core pedagogical principles: learning variability and math mindsets.

1. *Learning variability.* *ShaIA* promotes inclusive teaching practices aligned with the prioritized learning objectives of the national curriculum. A central feature of the platform is its integration of the Learner Variability framework (leveraging Digital Promise’s Learner Variability Navigator), which tailors pedagogical recommendations based on models capturing students’ diverse backgrounds, needs, and learning profiles.

Teachers incorporate learning variability into *ShaIA* through a three-step process. First, they select a learning model from four available options: mathematics, language, adult learning, and 21st-century learners. While *ShaIA* provides an initial recommendation, teachers may modify the model to better align it with their instructional context and objectives. Second, teachers define a class profile by selecting relevant variability factors that characterize their students (e.g., presence of non-Spanish-speaking students, migrants, students with disabilities, or students affected by trauma, etc.). Based on this profile, *ShaIA* presents a list of 35 factors identified as relevant for mathematics learning.⁵ Teachers then select between 10 and 15 factors they consider most salient for their classroom. Third, teachers select inclusion strategies. For each chosen factor, *ShaIA* suggests evidence-based pedagogical strategies that enable teachers to address classroom

⁵The factors are grouped into four domains: personal background; socio-emotional skills; cognition; and mathematics. The full list is presented in Appendix Table A.1.

diversity simultaneously.⁶ These strategies are automatically incorporated into the generation of lesson plans, activities, and instructional resources, ensuring that the platform's outputs are aligned with the actual characteristics of the class.

2. *Math mindsets.* *ShaIA*'s suggested pedagogical strategies are grounded in the Math Mindsets approach, which draws on robust recent evidence on effective mathematics teaching (Boaler and Dweck, 2016). This approach emphasizes understanding, creativity, and confidence, and is rooted in the concept of a growth mindset (Dweck, 2012) applied specifically to mathematics learning.

Building on this framework, *ShaIA* offers tools for lesson planning involving activity design, project-based learning, and development of formative assessments and grading rubrics. Its recommendations are informed by the selected learner variability factors and structured around five pedagogical tools of the math mindsets framework: mathematical experience, worksheets, number talks, pattern talks, and feedback practice. These tools are designed to strengthen mathematical communication, reasoning, problem solving, and modeling, while also supporting procedural fluency as an outcome of numerical flexibility and algebraic thinking rather than through isolated or mechanical practice.⁷

Overall, *ShaIA* is designed to complement rather than replace teachers. By automating lesson planning and reducing its preparation time, the platform aims to increase teacher engagement and promote higher-value activities such as instructional reflection and individualized student support. By moving beyond a one-size-fits-all approach, *ShaIA* seeks to enable teachers, particularly those with limited access to pedagogical resources, to adapt instructional strategies to classrooms with heterogeneous learning profiles.

Training on the use of *ShaIA*. Effective use of the platform requires that teachers understand its functionality and perceive its pedagogical value. To promote adoption and sustained use, *Mentú* and the SED implemented a structured training program for treated teachers. The program consisted of three school visits, three in-person workshops, and three one-hour webinars. The workshops were intended to familiarize teachers with the platform by modeling teaching strategies using *ShaIA*. The visits were intended to gather feedback and to review the teachers' work plan. Finally, the webinars were designed to reinforce *ShaIA*'s instructional applications, and support its integration into classroom practice.⁸

⁶Examples include making students' ideas visible, using multiple representations, anticipating diverse solution strategies, and fostering collaboration.

⁷Appendix A.2 provides a detailed discussion of the pedagogical approach.

⁸Appendix A.2 provide more details on the available resources offered to treated schools.

4 Research Design

4.1 Methodology

We evaluate the effects of providing teachers with AI support by offering school-level access to *ShaIA*. Treatment was assigned among 158 eligible schools in Bogotá, Colombia. Due to budget constraints, the agreement with the SED stipulated that teachers in lower secondary (grades 6th to 9th) in 58 schools would receive access to the platform. However, participation was voluntary and depended on each school principal’s willingness to join the evaluation. Schools originally assigned to treatment could therefore opt out. If they did, they neither received access to *ShaIA* nor participated in the experiment.

Randomization: Because of the risk of non-participation, we implemented a block-randomization algorithm in which schools were grouped into blocks and treatment was assigned within each block. This ensured that, even if a school assigned to treatment chose not to participate in a given block, the randomization would remain valid across the other blocks of the experiment. The randomization proceeded in three stages:

1. For each eligible school, we created a covariate index defined as the average of several school-level characteristics.⁹
2. Schools were then ranked by the covariate index and grouped into blocks. More similar schools were assigned to the same block.
3. Treatment was randomized within each block, yielding a treatment group of 70 schools and a control group of 88.

Panel A) of Table 2 summarizes the random assignment process, which occurred in two rounds. In the first round (described in columns (1) to (4) of panel A), 58 schools were offered treatment from a pool of 98 eligible schools grouped into 40 blocks. In this round, blocks consisted of two or three schools where one control school was assigned. In the second round (described in columns (5) to (8) of panel A), 12 additional schools were offered treatment from an eligible pool of 70 (that were not part of the first round). Here, blocks were formed by grouping five schools and randomly assigning one to the treatment group.

⁹All covariates were standardized and then averaged to construct a one-dimensional index. The covariates included: the total number of students in grades 6–9; indicators for rural location, morning session, afternoon session, and full-day provision; the share of low-income students (measured by the social stratum of students’ residences); average student age; the share of migrant students, students with disabilities, and female students; the total number of classrooms in grades 6–9; average scores on the high school exit exam in 2019 and 2020, including average mathematics scores; the total number of secondary school teachers and mathematics teachers; and the shares of contract teachers, female teachers, newly arrived teachers, secondary school teachers, and mathematics teachers.

Implementation: Not all schools who were offered the treatment decided to participate, affecting the implementation of the treatment. This procedure is described in Panel B) of Table 2. Of the 58 schools offered treatment in the first round, 45 accepted and 13 declined. Due to slower-than-anticipated take-up, four schools initially assigned to the control group were reassigned to treatment. As a result, 49 schools received treatment in the first round, but randomization was compromised in 13 blocks, involving 11 treated schools. In the remaining 26 blocks (comprising 34 treated and 26 control schools) random assignment proceeded as originally planned. In the second round, 12 schools were offered treatment. Two declined and were replaced by four schools originally assigned to the control group, invalidating three additional blocks. The second round therefore includes six valid blocks, consisting of six treated and 24 control schools.

These implementation patterns generate three mutually exclusive groups of schools, defined by their original block assignment: (i) blocks in which randomization proceeded without deviations (90 of 158 schools, 57%); (ii) blocks in which at least one school assigned to treatment declined participation (41 schools, 26%); and (iii) blocks in which control schools were reassigned to treatment (27 schools, 17%). These groups are described in the rows of Table 2. Our main estimation sample, referred to as "pure sample", consists of schools in blocks where randomization was not affected by take-up decisions. This sample includes teachers and students from 40 treated and 50 control schools for which treatment assignment was implemented as designed.

In addition to incomplete take-up, the intervention was not rolled out simultaneously across treated schools, introducing variation in treatment exposure. Due to capacity constraints, access to *ShaIA* and the associated workshops was delivered at different points during the academic year. We collect detailed information on the timing and duration of exposure to the platform, as well as on workshop participation, and use this variation to explore heterogeneity in treatment intensity and timing (see Section 5.2).

Teacher Participation: Participation in the experiment was voluntary. Teachers in some schools were offered access to the intervention but could choose not to participate. As a result, within treated schools we may observe both treated and untreated teachers. This does not invalidate our empirical strategy. However, it requires us to estimate both intent-to-treat and treatment-on-the-treated effects. Further details are provided in Section 5.

4.2 Data Collection

Figure 1 outlines the implementation of the evaluation, which will take place during the second semester of 2025 and throughout 2026. Randomization and treatment

assignment were conducted in August and September 2025. The 60 treated schools will receive access to ShaIA beginning in the second semester of 2025 and continuing through 2026. The evaluation is expected to conclude at the end of the academic year, in December 2026.

Students will be evaluated and teachers will be surveyed at two points in time. Baseline measurements were collected during August 2025. Due to budget constraints, baseline data were collected in a random sample of 69 schools (49 treated and 20 control). Endline measures will be collected at the end of the academic year, before November 2026.¹⁰ By that time, treated teachers will have had at least one full year of exposure to the platform. Because of similar budget limitations, a random sample of treated and control schools will be included in the endline. Current funding supports data collection in 69 schools, involving approximately 6,000 students, although this sample size may increase if additional resources are secured before endline data collection begins.¹¹

Teacher data will include sociodemographic characteristics, information on teaching practices, their use of digital technologies, and their perceptions of AI. Importantly, at baseline we will ask teachers whether they know about ShaIA, but no additional information about the platform will be collected, as it is expected that neither teachers in treated nor control schools are familiar with the program at that stage. At endline, we will collect detailed information on teachers' perceptions and experiences with ShaIA. We expect teachers in the treatment group to report substantially greater familiarity with the platform, whereas most teachers in the control group will likely remain unaware of its existence.

Student data will include basic demographic characteristics, information on the use of technologies, educational expectations, and math test scores from a paper-based assessment. This assessment is designed by *Im-prove*, a local expert company specializing in the development of grade-appropriate tests for Colombian students. These tests are pre-approved by the SED and aligned with the curriculum implemented in Bogotá's public schools. All test items were pre-tested in an alternative setting to ensure their validity and suitability for the local student population. Importantly, student data will be collected irrespective of teacher participation (i.e., observing students whose teachers were and were not treated), allowing us to conduct a proper comparison between treatment and control groups.¹²

¹⁰The date of the final date of data collection will depend on the SED's willingness to extend the intervention either through the end of the academic year or through the end of the first semester. This implies that endline could occur either around July or November 2026.

¹¹We also plan to include a long-run test score measure using the high school exit exam. Given that randomization was conducted at the school level, this allows us to estimate longer-run intent-to-treat effects using administrative, publicly available data. These long-run results will be included in a separate paper.

¹²More information about Im-prove exams can be found at: <https://www.improvecolombia.com/>.

4.3 Outcomes

Outcomes for both treatment and control groups will be constructed using the information gathered during data collection. Two main categories of outcomes are relevant for our study: teacher-level outcomes and student-level outcomes.

- *Teachers:* Teachers represent the first step in the successful adoption of AI in the classroom. We expect the intervention to influence teaching strategies, which may in turn affect student learning and reduce variability in performance. The following outcomes will be collected to quantify potential changes in teaching practices:

1. **Time spent preparing class:** At baseline and endline, teachers will report the amount of time spent preparing their most recently taught class.
2. **Share of class time devoted to different activities:** Teachers will report the proportion of class time allocated to lecturing, practical exercises, group work, discussions, use of technology, and assessments.
3. **Math Teaching Index:** A principal component index capturing math teaching aptitude. It is constructed from the following items, each measured on a 1–5 scale:
 - I feel confident teaching math concepts.
 - My students actively participate in math classes.
 - I use real-world examples to teach math.
 - My math classes encourage problem-solving skills.
 - Students collaborate with their peers during math activities.
 - Technology enhances student learning in my math lessons.
 - I have enough resources to teach mathematics effectively.
 - My students can explain their mathematical reasoning clearly.
 - It is beneficial for the class when students share different perspectives.
 - I teach my students that anyone can become more intelligent at math.
 - I provide students with individual feedback.
4. **AI Perception Index:** A principal component index measuring teachers' perceptions and use of AI in the classroom. It is based on the following items, each on a 1–5 scale:
 - I understand how AI can be used to support math instruction.
 - I feel confident using AI to generate instructional material.
 - I have received adequate training on how to use AI effectively.

- I use AI-generated content to explain math concepts.
 - I evaluate AI-generated materials to ensure they are appropriate for students with differentiated learning.
 - AI helps me identify and address gaps in student learning.
 - I use AI-driven analytics to track student progress and adjust my teaching.
 - I have access to AI-generated lesson plans, activities, or instructional sequences aligned with prioritized learning objectives.
 - I modify AI-generated resources to better suit the needs of my students.
 - My educational institution provides guidance on choosing and adapting AI-generated resources.
 - AI has improved the efficiency and quality of my lesson planning.
- *Students*: Our primary student-level outcome consists of test score measures designed and collected by *Im-prove*. The assessment includes 18 multiple-choice questions, each with four response options, and is expected to be completed in 50 minutes.¹³ The outcome corresponds to the share of correct answers, standardized relative to the mean and standard deviation of test-takers in each round. Although baseline and endline questions differ, both assessments are aligned with the curriculum and the student’s grade level.

5 Empirical Strategy

Due to the random nature of the treatment assignment, we can estimate the effect of providing teachers with AI support by comparing outcomes for teachers and students in treated schools with those in control schools. Because treatment was assigned using block randomization, these comparisons must be made within blocks; therefore, we condition on block indicators. Formally, we estimate:

$$y_{i,j \in b} = \alpha + \beta_{ITT} T_j + \gamma_b + \varepsilon_{i,j \in b}, \quad (1)$$

where $y_{i,j \in b}$ denotes the outcome for teacher or student i in school j within block b . The variable T_j is an indicator equal to one if school j was assigned to the treatment group and zero otherwise, and γ_b represents block fixed effects. Standard errors are clustered at the school level.

Some mathematics teachers in treated schools may choose not to participate in the intervention. As a result, some units assigned to treatment may remain untreated,

¹³Testing was conducted in one-hour sessions, but 10 minutes were used to organize the classroom and give instructions, and deliver the material.

implying that Equation 2 identifies an Intention-to-Treat (ITT) effect. To account for imperfect compliance, we additionally define an indicator variable, D_i , equal to one if teacher i adopts the technology. We then instrument treatment adoption (D_i) with treatment assignment (T_j) to estimate the Treatment-on-the-Treated (TOT) effect. Formally, this approach entails estimating the following two equations:

$$\begin{aligned} D_{i,j \in b} &= \gamma_0 + \gamma_1 T_j + \gamma_b + \epsilon_{i,j \in b}, \\ y_{i,j \in b} &= \delta_0 + \beta_{TOT} \hat{D}_i + \gamma_b + u_{i,j \in b}. \end{aligned} \quad (2)$$

5.1 Internal Validity

The internal validity of the experiment relies on treatment being assigned orthogonally to baseline characteristics. To assess this, we conduct balance tests using school-level administrative data (Table 3) and baseline survey data (Table 4). For each table, we present results for the full sample, for the sample excluding the control schools that were mistakenly treated, as well as for the pure sample that excludes blocks with non-participants. We report both unadjusted p-values and p-values adjusted for block fixed effects. Despite the implementation issues, the evidence indicates that the randomization was successful: treatment is orthogonal to observable characteristics, and the absence of statistically significant differences across groups strongly supports this conclusion.

5.2 Heterogeneity

Treatment effects of educational interventions are known to be highly heterogeneous (Buhl-Wiggers et al., 2024). This makes it essential to study effect heterogeneity when evaluating the impact of ShaIA. We approach this in two ways.

First, the context of Bogotá suggests substantial pre-existing gaps in math achievement by gender and migrant status. Prior evidence shows that girls in Colombia score significantly lower in math, a gap largely explained by within-school factors (Muñoz, 2018). Likewise, migrant students tend to learn less in non-adapted environments, and lower learning rates for migrant populations have been documented in Colombia as well. Table 5 confirms these patterns in our baseline sample: both girls and migrant students exhibit lower math achievement prior to the intervention. These patterns provide theoretical grounds to expect larger treatment effects for these groups, given that ShaIA aims specifically to address within-classroom heterogeneity. Therefore, we include these two heterogeneity dimensions in our pre-analysis plan as pre-specified heterogeneous effects that test our original hypothesis.

In addition, given the nature of the treatment, we allow for the possibility that treatment effects may vary across teacher characteristics and levels of treatment exposure. Accordingly, we will examine treatment effect heterogeneity by teacher type—specifically age, baseline AI knowledge, and gender—as well as by treatment exposure (time exposed to *ShaIA* and timing of the workshops).

Second, following [Buhl-Wiggers et al. \(2024\)](#), we will formally test the null hypothesis that the standard deviation of treatment effects is equal to zero. Rejecting this null would indicate heterogeneous impacts, in which case we will estimate Conditional Average Treatment Effects (CATEs) based on pre-specified characteristics, following the framework of [Wager and Athey \(2018\)](#). This data-driven analysis will help identify the subgroups that benefit the most from the intervention. Our hypothesis is that girls and migrant students are likely to experience the largest gains, as the intervention is designed to reduce within-class learning gaps.

5.3 Statistical Power

Our intervention is a randomized trial clustered at the school level, as schools constitute the unit of randomization. Accordingly, our power calculations account for clustering and assume that sampling occurs at the classroom level. For these calculations, we assume 80% power, an intra-cluster correlation coefficient of 0.1, a 5% significance level, two sampled per school, and 30 students per teacher.

Current funding allows us to collect data from approximately 6,000 students across 40 treatment schools and 50 control schools.¹⁴ Under this design, we have sufficient power to detect an overall average treatment effect on mathematics achievement of 0.14 standard deviations (SD).

Assuming that 50% of students are female and 10% are migrant yields minimum detectable effects of 0.2 SD for girls and 0.28 SD for migrant students. For reference, a 4.5-month technology-aided after-school instruction program in India delivered an impact of 0.37 SD ([Muralidharan et al., 2019](#)).

5.4 Implementation Costs

The total implementation cost of the program is \$1,096 USD per teacher, of which \$264.12 covers platform licenses and access, and \$832.50 covers additional support.¹⁵ On average, each math teacher in the sample teaches 120 students, resulting in a total

¹⁴Efforts are ongoing to increase the sample size for the endline collection. If successful, additional classrooms may be included, with potential oversampling of migrant students.

¹⁵The total cost in Colombian pesos is equivalent to 4,442,891, which we convert to USD using the average 2025 exchange rate of 4,051.4636.

implementation cost of \$9.14 USD per student—\$2.20 for licenses and \$6.94 for support.

Figures and Tables

Figure 1: Timeline of Intervention and Data Collection

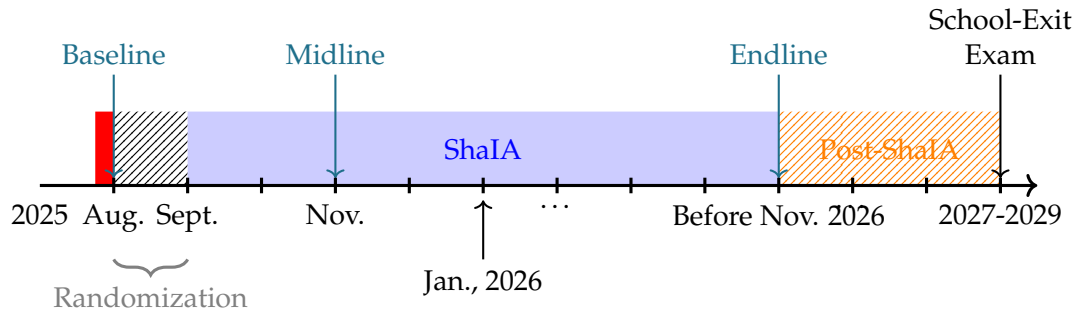


Table 1: Test Score Gaps in High School Exit Exam

	Income			Gender			Migration		
	Public (1)	Private (2)	Gap (3)	Women (4)	Men (5)	Gap (6)	Migrant (7)	Local (8)	Gap (9)
Overall	-0.138	0.437	-0.574	-0.064	0.075	-0.139	-0.174	0.002	-0.176
Math	-0.112	0.356	-0.468	-0.113	0.134	-0.247	-0.220	0.003	-0.223

Note: Data come from the 2022 edition of *Saber 11*. Test scores are normalized to have mean zero and standard deviation of one.

Table 2: School Randomization and Implementation

	Round 1					Round 2					Total		Overall (13)
	Blocks (1)	Non-Part. (2)	Treated (3)	Control (4)	Total (5)	Blocks (6)	Non-Part. (7)	Treated (8)	Control (9)	Total (10)	Treated (11)	Control (12)	
<i>A) Original Randomization</i>													
Pure Sample	26		34	26	60	6		6	24	30	40	50	90
With Non-Participant	9		17	9	26	3		3	12	15	20	21	41
With Switcher	5		7	5	12	3		3	12	15	10	17	27
Total	40		58	40	98	12		12	48	60	70	88	158
<i>B) Implementation</i>													
Pure Sample	26		34	26	60	6		6	24	30	40	50	90
With Non-Participant	9	12	5	9	26	3	3		12	15	5	21	41
With Switcher	5	5	6	1	12	3	2	4	9	15	10	10	27
Total	40	17	45	36	98	12	5	10	45	60	55	81	158

Table 3: Balance Across Experimental Groups at Baseline –School Level Outcomes

	Universe						Dropping Switchers						Dropping Declined					
	Observations		Average		P-Value		Observations		Average		P-Value		Observations		Average		P-Value	
	Treat.	Con.	Treat.	Con.	Unadj.	Adj.	Treat.	Con.	Treat.	Con.	Unadj.	Adj.	Treat.	Con.	Treat.	Con.	Unadj.	Adj.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
<i>A) Teachers</i>																		
Number of Teachers	55	103	86.382	85.097	0.845	0.730	45	86	84.733	87.256	0.706	0.932	40	50	84.675	91.260	0.452	0.928
Number of Teachers Sec. School	55	103	42.836	41.728	0.738	0.996	45	86	42.022	42.651	0.856	0.552	40	50	42.250	45.140	0.523	0.511
Number of Math Teachers	55	103	5.873	5.369	0.341	0.589	45	86	5.889	5.512	0.519	0.186	40	50	5.975	5.800	0.810	0.146
Number of Math Teachers Sec. School	55	103	5.745	5.311	0.413	0.698	45	86	5.800	5.453	0.554	0.220	40	50	5.875	5.740	0.853	0.191
Permanent (%)	55	103	0.854	0.875	0.182	0.127	45	86	0.842	0.877	0.054	0.039*	40	50	0.843	0.878	0.103	0.055
Female (%)	55	103	0.667	0.675	0.454	0.639	45	86	0.665	0.681	0.113	0.365	40	50	0.668	0.680	0.316	0.704
Hired Under New Regime–2277 (%)	55	103	0.207	0.249	0.045*	0.024*	45	86	0.188	0.257	0.002*	0.002*	40	50	0.188	0.262	0.005*	0.002*
Secondary School (%)	55	103	0.496	0.489	0.503	0.530	45	86	0.494	0.488	0.596	0.423	40	50	0.496	0.494	0.866	0.395
Math (%)	55	103	0.066	0.063	0.310	0.375	45	86	0.067	0.064	0.398	0.230	40	50	0.067	0.064	0.514	0.225
Math Teachers in Sec. School (%)	55	103	0.130	0.127	0.634	0.603	45	86	0.133	0.128	0.486	0.274	40	50	0.133	0.127	0.456	0.306
Years Working in School	55	103	8.977	9.932	0.036*	0.018*	45	86	8.761	10.221	0.003*	0.003*	40	50	8.837	10.243	0.011*	0.006*
Age	55	103	48.122	48.935	0.155	0.055	45	86	47.854	49.162	0.033*	0.014*	40	50	48.004	49.465	0.033*	0.021*
<i>B) Students</i>																		
Number of Students	55	102	598.400	600.265	0.974	0.453	45	85	590.933	618.588	0.639	0.763	40	49	589.775	665.388	0.333	0.702
Number of Students in 6th Grade	55	102	163.055	163.990	0.954	0.332	45	85	159.289	168.400	0.569	0.507	40	49	158.425	178.735	0.338	0.555
Number of Students in 7th Grade	55	102	154.727	154.755	0.999	0.396	45	85	152.333	159.212	0.662	0.649	40	49	152.150	169.796	0.391	0.654
Number of Students in 8th Grade	55	102	144.364	146.490	0.881	0.400	45	85	141.978	151.071	0.533	0.716	40	49	141.925	165.694	0.225	0.550
Number of Students in 9th Grade	55	102	136.255	135.029	0.925	0.977	45	85	137.333	139.906	0.854	0.630	40	49	137.275	151.163	0.447	0.765
Students in Rural Schools (%)	55	102	0.018	0.010	0.686	0.997	45	85	0.022	0.012	0.679	0.997	40	49	0.025	0.000	0.319	0.428
Students in Morning Session (%)	55	102	0.408	0.389	0.718	0.318	45	85	0.435	0.399	0.547	0.452	40	49	0.406	0.374	0.649	0.688
Students in Afternoon Session (%)	55	102	0.283	0.297	0.748	0.801	45	85	0.276	0.306	0.526	0.672	40	49	0.270	0.302	0.569	0.766
Students in All-day Session (%)	55	102	0.309	0.314	0.944	0.578	45	85	0.289	0.295	0.941	0.759	40	49	0.325	0.324	0.993	0.908
Low-Income Students (%)	55	102	0.308	0.270	0.252	0.718	45	85	0.328	0.274	0.163	0.414	40	49	0.324	0.264	0.173	0.354
Age	55	102	14.511	14.458	0.184	0.399	45	85	14.515	14.461	0.231	0.526	40	49	14.507	14.479	0.649	0.816
Migrant Students (%)	55	102	0.118	0.121	0.731	0.859	45	85	0.118	0.118	0.948	0.752	40	49	0.117	0.109	0.487	0.639
Students with Disabilities (%)	55	102	0.034	0.039	0.313	0.192	45	85	0.033	0.040	0.314	0.137	40	49	0.033	0.044	0.284	0.183
Female Students (%)	55	102	0.481	0.499	0.042*	0.072	45	85	0.478	0.503	0.014*	0.048*	40	49	0.479	0.502	0.067	0.121
<i>C) Classrooms</i>																		
Number of Sections	55	102	12.455	12.343	0.893	0.752	45	85	12.378	12.847	0.599	0.855	40	49	12.450	13.714	0.276	0.565
Number of Sections in 6th Grade	55	102	3.255	3.294	0.861	0.363	45	85	3.200	3.412	0.365	0.379	40	49	3.200	3.633	0.167	0.208
Number of Sections in 7th Grade	55	102	3.200	3.147	0.811	0.744	45	85	3.178	3.282	0.665	0.767	40	49	3.200	3.490	0.363	0.572
Number of Sections in 8th Grade	55	102	3.109	3.000	0.634	0.737	45	85	3.089	3.106	0.946	0.586	40	49	3.100	3.367	0.390	0.992
Number of Sections in 9th Grade	55	102	2.927	2.902	0.900	0.964	45	85	2.956	3.047	0.682	0.774	40	49	3.000	3.224	0.432	0.849
<i>C) Test Scores</i>																		
Average Saber 11 2019	50	96	0.085	0.109	0.567	0.384	41	79	0.062	0.094	0.466	0.165	38	46	0.066	0.089	0.655	0.164
Average Saber 11 2020	49	96	0.036	0.044	0.845	0.558	40	79	0.016	0.036	0.666	0.328	37	46	0.019	0.052	0.534	0.362
Math Saber 11 2019	50	96	0.076	0.094	0.664	0.230	41	79	0.055	0.080	0.564	0.058	38	46	0.060	0.072	0.830	0.061
Math Saber 11 2020	49	96	0.051	0.057	0.890	0.389	40	79	0.037	0.056	0.680	0.169	37	46	0.038	0.060	0.698	0.190
Missing Saber 11 (%)	55	103	0.091	0.068	0.621	0.313	45	86	0.089	0.081	0.886	0.524	40	50	0.050	0.080	0.567	0.771

Table 4: Balance Across Experimental Groups at Baseline

	Universe						Dropping Switchers						Dropping Declined					
	Observations		Average		P-Value		Observations		Average		P-Value		Observations		Average		P-Value	
	Treat.	Con.	Treat.	Con.	Unadj.	Adj.	Treat.	Con.	Treat.	Con.	Unadj.	Adj.	Treat.	Con.	Treat.	Con.	Unadj.	Adj.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
<i>A) Students</i>																		
Baseline Test Score	4,080	1,655	31.80	32.10	0.78	0.30	3,515	1,571	31.54	31.99	0.69	0.36	3,115	1,048	31.58	32.68	0.44	0.32
Baseline Test Score (σ)	4,080	1,655	-0.02	-0.00	0.78	0.30	3,515	1,571	-0.04	-0.01	0.69	0.36	3,115	1,048	-0.03	0.04	0.44	0.32
Females	4,060	1,650	0.48	0.49	0.55	0.40	3,497	1,566	0.48	0.49	0.56	0.36	3,099	1,044	0.48	0.48	0.85	0.98
Migrants	4,080	1,655	0.32	0.32	0.99	0.43	3,515	1,571	0.32	0.32	0.99	0.41	3,115	1,048	0.31	0.29	0.63	0.61
International Migrants	4,080	1,655	0.09	0.09	0.73	0.26	3,515	1,571	0.09	0.10	0.66	0.23	3,115	1,048	0.09	0.08	0.85	0.45
Wealth PC Index	4,080	1,655	0.02	-0.05	0.46	0.15	3,515	1,571	0.02	-0.07	0.33	0.10	3,115	1,048	0.03	-0.01	0.72	0.23
Use of AI	4,079	1,655	0.84	0.86	0.36	0.03*	3,514	1,571	0.84	0.86	0.34	0.04*	3,114	1,048	0.84	0.85	0.95	0.31
Advance Knowledge of AI	4,080	1,655	0.30	0.32	0.25	0.11	3,515	1,571	0.30	0.32	0.25	0.11	3,115	1,048	0.30	0.33	0.24	0.02*
Math Learning Perceptions	4,031	1,635	0.03	-0.07	0.17	0.65	3,471	1,551	0.03	-0.06	0.20	0.68	3,071	1,036	0.05	-0.03	0.16	0.06
Math Class Perceptions	4,030	1,635	0.01	-0.02	0.51	0.82	3,471	1,551	0.01	-0.02	0.60	0.89	3,075	1,032	0.02	0.02	1.00	0.90
Teacher Perceptions	3,968	1,606	0.01	-0.04	0.54	1.00	3,417	1,522	0.01	-0.03	0.60	0.91	3,030	1,014	0.05	0.03	0.79	0.91
Expected to Complete College	4,080	1,655	0.82	0.81	0.85	0.36	3,515	1,571	0.81	0.81	0.76	0.36	3,115	1,048	0.82	0.81	0.68	0.63
Expected to Continue in STEM	4,080	1,655	0.24	0.22	0.16	0.07	3,515	1,571	0.24	0.22	0.08	0.02*	3,115	1,048	0.24	0.21	0.09	0.02*
AI can be helpful in the future	4,080	1,655	0.34	0.32	0.30	0.75	3,515	1,571	0.35	0.32	0.21	0.53	3,115	1,048	0.35	0.32	0.07	0.03*
<i>B) Teacher Characteristics</i>																		
Female	167	101	0.47	0.40	0.24	0.16	158	97	0.47	0.39	0.19	0.08	135	67	0.45	0.45	0.96	0.63
Age	167	101	45.74	46.37	0.60	0.18	158	97	45.63	46.37	0.55	0.06	135	67	46.06	46.43	0.80	0.37
High School or Less	167	101	0.05	0.08	0.25	0.27	158	97	0.05	0.08	0.26	0.27	135	67	0.06	0.04	0.63	0.73
University Degree	167	101	0.09	0.05	0.19	0.23	158	97	0.09	0.05	0.18	0.23	135	67	0.07	0.06	0.67	0.64
Postgraduate	167	101	0.86	0.87	0.79	1.00	158	97	0.85	0.87	0.74	1.00	135	67	0.87	0.90	0.46	0.52
Teaching Experience	167	101	18.69	19.93	0.21	0.12	158	97	18.58	20.01	0.16	0.06	135	67	19.10	19.63	0.64	0.51
Work Experience	167	101	22.18	23.99	0.08	0.01*	158	97	22.04	23.95	0.07	0.00*	135	67	22.41	24.01	0.16	0.08
<i>C) Use of Digital Technologies</i>																		
Digital Resources at home (Index)	159	99	-0.00	0.00	0.99	0.47	150	95	-0.02	-0.01	0.93	0.41	128	66	-0.01	-0.05	0.83	0.40
Digital Resources at work (Index)	155	93	0.14	-0.23	0.08	0.11	147	89	0.13	-0.24	0.09	0.14	125	62	0.15	-0.22	0.18	0.16
First computer use age > 20	167	101	0.29	0.27	0.71	0.10	158	97	0.29	0.27	0.68	0.03*	135	67	0.30	0.25	0.41	0.23
First internet use age > 20	167	101	0.42	0.44	0.78	0.13	158	97	0.42	0.43	0.80	0.05	135	67	0.43	0.43	0.96	0.12
Internet > 2h per day	167	101	0.68	0.67	0.89	0.45	158	97	0.68	0.66	0.74	0.32	135	67	0.70	0.61	0.36	0.05
Internet at School > 2h per day	167	101	0.49	0.40	0.21	0.19	158	97	0.49	0.38	0.14	0.17	135	67	0.50	0.34	0.04*	0.06
Uses AI at Work	167	101	0.68	0.76	0.23	0.78	158	97	0.68	0.75	0.32	0.48	135	67	0.68	0.75	0.43	0.79
Knowledge of AI	167	101	0.83	0.87	0.32	0.63	158	97	0.84	0.87	0.45	0.33	135	67	0.82	0.87	0.37	0.78
Knowledge of AI to teach	167	101	0.74	0.83	0.08	0.72	158	97	0.73	0.82	0.08	0.72	135	67	0.72	0.82	0.14	0.65
Uses AI to Teach	167	101	0.41	0.51	0.12	0.47	158	97	0.41	0.51	0.19	0.84	135	67	0.39	0.52	0.04*	0.05
Digital Use (Index)	154	86	0.07	-0.13	0.42	0.00*	146	83	0.02	-0.17	0.43	0.00*	125	59	0.02	-0.09	0.64	0.01*
Digital Use In-Class (Index)	161	95	0.03	-0.04	0.68	0.13	152	92	-0.01	-0.06	0.81	0.14	129	64	0.08	0.04	0.85	0.06
Perc. Digital Use in Class (Index)	153	94	-0.14	0.22	0.09	0.46	145	90	-0.13	0.30	0.04*	0.35	124	64	-0.20	0.24	0.03*	0.09
<i>D) Teacher Practices</i>																		
Last class prepared > 1h	167	101	0.33	0.34	0.89	0.19	158	97	0.32	0.35	0.50	0.32	135	67	0.30	0.33	0.63	0.73
Teacher Aptitude (Index)	161	92	-0.07	0.12	0.42	0.79	153	88	-0.05	0.18	0.34	0.70	131	60	-0.09	0.28	0.23	0.95
Class time: Lecture (%)	167	101	25.26	26.93	0.28	0.04*	158	97	25.53	27.32	0.25	0.03*	135	67	25.32	27.76	0.22	0.03*
Class time: Practical Exercises (%)	167	101	23.05	22.15	0.51	0.06	158	97	23.13	22.29	0.56	0.06	135	67	23.33	22.34	0.57	0.08
Class time: Group Work (%)	167	101	17.57	18.25	0.58	0.36	158	97	17.46	18.07	0.63	0.47	135	67	17.66	18.03	0.82	0.53
Class time: Discussion (%)	167	101	11.58	10.71	0.25	0.26	158	97	11.51	10.64	0.27	0.31	135	67	11.47	10.63	0.40	0.52
Class time: Use of Technology (%)	167	101	5.96	5.67	0.76	0.99	158	97	5.89	5.24	0.45	0.78	135	67	5.79	4.75	0.29	0.77
Class time: Assessments (%)	167	101	10.87	10.87	1.00	0.56	158	97	10.84	11.01	0.85	0.36	135	67	10.76	11.13	0.68	0.47
Class time: Other (%)	167	101	5.70	5.42	0.70	0.86	158	97	5.63	5.43	0.80	0.86	135	67	5.67	5.36	0.72	0.78
Teacher Own Perceptions (Index)	158	94	-0.10	0.17	0.29	0.13	149	90	-0.13	0.17	0.27	0.19	129	63	-0.17	0.26	0.20	0.11
<i>E) Perceptions in the Use of AI</i>																		
Perceptions on AI to teach (Index)	141	89	-0.12	0.18	0.33	0.03*	134	85	-0.04	0.20	0.44	0.04*	114	59	-0.04	0.18	0.57	0.08
AI is useful to teach	167	101	0.60	0.45	0.02*	0.00*	158	97	0.58	0.42	0.01*	0.00*	135	67	0.60	0.46	0.03*	0.00*
AI helps: Class prep.	167	101	0.98	0.93	0.04*	0.00*	158	97	0.98	0.93	0.04*	0.00*	135	67	0.98	0.93	0.10	0.03*
AI helps: Resources	167	101	0.97	0.91	0.07	0.01*	158	97	0.97	0.92	0.13	0.01*	135	67	0.96	0.91	0.23	0.07
AI helps: Track Learning	167	101	0.80	0.66	0.06	0.50	158	97	0.79	0.65	0.04*	0.50	135	67	0.82	0.66	0.05	0.10
AI helps: Personalized Teaching	167	101	0.92	0.83	0.13	0.23	158	97	0.91	0.82	0.13	0.23	135	67	0.90	0.82	0.25	0.37
Last Year: Seminar on AI (%)	167	101	0.21	0.26	0.30	0.34	158	97	0.21	0.26	0.31	0.29	135	67	0.22	0.27	0.43	0.42
Last Year: Course on AI (%)	167	101	0.23	0.29	0.21	0.08	158	97	0.23	0.27	0.40	0.27	135	67	0.23	0.25	0.59	0.09
Last Year: Use an AI Tool (%)	167	101	0.64	0.66	0.73	0.33	158	97	0.64	0.65	0.88	0.21	135	67	0.64	0.66	0.88	0.33
Last Year: Use AI Tool to Teach (%)	167	101	0.46	0.50	0.42	0.32	158	97	0.46	0.49	0.54	0.16	135	67	0.45	0.49	0.56	0.51
Last Year: Research on AI (%)	167	101	0.49	0.47	0.76	0.09	158	97	0.47	0.47	0.99	0.16	135	67	0.49	0.46	0.73	0.17
Last Year: Discussed AI (%)	167	101	0.51	0.52	0.81	0.79	158	97	0.49	0.52	0.74	0.66	135	67	0.50	0.49	0.96	0.86
Last Year: Other AI Training (%)	167	101	0.16	0.19	0.63	0.30	158	97	0.16	0.20	0.50	0.30	135	67	0.16	0.19	0.66	0.29
¿Conoce la herramienta ShalA?	165	99	0.07	0.03	0.11	0.54	156	95	0.06	0.03	0.22	0.90	133	66	0.08	0.05	0.38	0.90

Table 5: Baseline Test Score Differences by Gender and Migrant Status

	Migrant Status						Gender		
	Obs	Locals	Migrants		Differences		Female	Male	Difference (7) vs (8)
			Internal	External	(2) vs (3)	(2) vs (4)			
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	
Grade 6	1,596	0.255	0.265	0.112	0.018	-0.143*	0.248	0.245	-0.002
Grade 7	1,425	0.421	0.152	0.196	-0.266***	-0.213**	0.382	0.339	-0.043
Grade 8	1,475	-0.384	-0.415	-0.439	-0.030	-0.056	-0.369	-0.417	-0.047
Grade 9	1,239	-0.297	-0.413	-0.508	-0.115*	-0.211***	-0.238	-0.424	-0.186***
Overall	5,735	0.014	-0.103	-0.120	-0.095**	-0.154***	0.020	-0.048	-0.068**

Note: Columns (4), (5), and (8) present differences computed using a regression of test scores on characteristic dummies, including randomization block fixed effects. Standard errors are clustered at the school level. *** p<0.01, ** p<0.05, * p<0.1

References

- Andrabi, T., Baron, J., Macdonald, I., and Qureshi, Z. (2025). Free to choose: Introducing technology for foundational learning in pakistan. Mimeo.
- Angrist, J. and Lavy, V. (2002). New evidence on classroom computers and pupil learning*. *The Economic Journal*, 112(482):735–765.
- Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., and Schady, N. (2016). Teacher Quality and Learning Outcomes in Kindergarten. *The Quarterly Journal of Economics*, 131(3):1415–1453.
- Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., and Walton, M. (2017). From proof of concept to scalable policies: Challenges and solutions, with an application. *Journal of Economic Perspectives*, 31(4):73–102.
- Banerjee, A. V., Banerji, R., Duflo, E., Glennerster, R., and Khemani, S. (2010). Pitfalls of participatory programs: Evidence from a randomized evaluation in education in india. *American Economic Journal: Economic Policy*, 2(1):1–30.
- Banerjee, A. V., Cole, S., Duflo, E., and Linden, L. (2007). Remedying Education: Evidence from Two Randomized Experiments in India*. *The Quarterly Journal of Economics*, 122(3):1235–1264.
- Barrow, L., Markman, L., and Rouse, C. E. (2009). Technology’s edge: The educational benefits of computer-aided instruction. *American Economic Journal: Economic Policy*, 1(1):52–74.
- Bassi, M., Busso, M., and Muñoz, J. S. (2015). Enrollment, graduation, and dropout rates in latin america: Is the glass half empty or half full? *Economía*, 16(1):113–156.
- Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., and Mariman, R. (2024). Generative ai can harm learning. SSRN Scholarly Paper No. 4895486. Accessed: 2026-01-08.
- Berlinski, S. and Busso, M. (2017). Challenges in educational reform: An experiment on active learning in mathematics. *Economics Letters*, 156:172–175.
- Beuermann, D. W., Cristia, J., Cueto, S., Malamud, O., and Cruz-Aguayo, Y. (2015). One laptop per child at home: Short-term impacts from a randomized experiment in peru. *American Economic Journal: Applied Economics*, 7(2):53–80.
- Blimpo, M. P., Gajigo, O., Tomita, R., Owusu, S., and Xu, Y. (2020). Technology in the classroom and learning in secondary schools. Policy research working paper; no. 9288, World Bank, Washington, DC. Accessed: 2026-01-08.
- Boaler, J. and Dweck, C. (2016 - 2016). *Mathematical mindsets : unleashing students’ potential through creative math, inspiring messages, and innovative teaching*. Jossey-Bass, San Francisco, California, 1st ed. edition.
- Büchel, K., Jakob, M., Kühnhanss, C., Steffen, D., and Brunetti, A. (2022). The relative effectiveness of teachers and learning software: Evidence from a field experiment in el salvador. *Journal of Labor Economics*, 40(3):737–777.

- Buhl-Wiggers, J., Kerwin, J. T., Muñoz-Morales, J., Smith, J., and Thornton, R. (2024). Some children left behind: Variation in the effects of an educational intervention. *Journal of Econometrics*, 243(1):105256.
- Busso, M., Montaña, S., Muñoz-Morales, J., and Pope, N. G. (2024). The unintended consequences of merit-based teacher selection: Evidence from a large-scale reform in Colombia. *Journal of Public Economics*, 239:105238.
- Chernozhukov, V., Demirer, M., Duflo, E., and Fernández-Val, I. (2022). Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Immunization in India. Working Paper 24678, National Bureau of Economic Research.
- Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., and Yagan, D. (2011). How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star. *The Quarterly Journal of Economics*, 126(4):1593–1660.
- Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American Economic Review*, 104(9):2633–79.
- Cristia, J., Ibararán, P., Cueto, S., Santiago, A., and Severín, E. (2017). Technology and child development: Evidence from the one laptop per child program. *American Economic Journal: Applied Economics*, 9(3):295–320.
- De Simone, M. E., Tibetti, F., Barron, M., Manolio, F., Mosuro, W., and Dikoru, E. (2025). From Chalkboards to Chatbots : Evaluating the Impact of Generative AI on Learning Outcomes in Nigeria. Policy research working paper, World Bank Group.
- Duflo, E., Dupas, P., and Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review*, 101(5):1739–74.
- Dweck, C. (2012). *Mindset: Changing The Way You think To Fulfil Your Potential*. Little, Brown Book Group.
- Escueta, M., Quan, V., Nickow, A. J., and Oreopoulos, P. (2017). Education technology: An evidence-based review. Working Paper 23744, National Bureau of Economic Research.
- Han, H., Kerwin, J. T., Muñoz-Morales, J., Smith, J., and Thornton, R. (2026). How many children left behind? a meta-analysis of treatment effect heterogeneity in developing-country education programs. Mimeo.
- Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review*, 30(3):466–479.
- Hanushek, E. A. and Rivkin, S. G. (2006). Chapter 18 teacher quality. In Hanushek, E. and Welch, F., editors, *Handbook of the Economics of Education*, volume 2, pages 1051–1078. Elsevier.
- Hanushek, E. A. and Rivkin, S. G. (2012). The distribution of teacher quality and implications for policy. *Annual Review of Economics*, 4(1):131–157.

- Henkel, O., Horne-Robinson, H., Kozhakhmetova, N., and Lee, A. (2024). Effective and scalable math support: Experimental evidence on the impact of an ai-math tutor in ghana. In *Artificial Intelligence in Education: Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, Communications in Computer and Information Science, pages 373–381. Springer Nature Switzerland. Accessed: 2026-01-08.
- Icfes (2024). Programa para la evaluación internacional de alumnos (pisa). informe nacional de resultados para colombia 2022. Technical report, ICFES.
- Jackson, C. K. (2018). What do test scores miss? the importance of teacher effects on non-test score outcomes. *Journal of Political Economy*, 126(5):2072–2107.
- Jackson, K. and Makarin, A. (2018). Can online off-the-shelf lessons improve student outcomes? evidence from a field experiment. *American Economic Journal: Economic Policy*, 10(3):226–54.
- Lehmann, M., Cornelius, P. B., and Sting, F. J. (2024). Ai meets the classroom: When does chatgpt harm learning? arXiv preprint arXiv:2409.09047v1. Accessed: 2026-01-08.
- Lopez, C., Molina, E., and Zárate, R. A. (2025). Ai in the classroom: Evaluating the impact of teacher training on practices and student outcomes. Mimeo.
- Ma, Y., Fairlie, R., Loyalka, P., and Rozelle, S. (2024). Isolating the “tech” from edtech: Experimental evidence on computer-assisted learning in china. *Economic Development and Cultural Change*, 72(4):1923–1962.
- Melo-Becerra, L. A., Ramos-Forero, J. E., Rodríguez Arenas, J. L., and Zára, H. M. (2021). Efecto de la pandemia sobre el sistema educativo: El caso de colombia. Borradores de Economía 1179, Banco de la República de Colombia.
- Muralidharan, K., Singh, A., and Ganimian, A. J. (2019). Disrupting education? experimental evidence on technology-aided instruction in india. *American Economic Review*, 109(4):1426–60.
- Muñoz, J. S. (2018). The economics behind the math gender gap: Colombian evidence on the role of sample selection. *Journal of Development Economics*, 135:368–391.
- Oreopoulos, P., Gibbs, C., Jensen, M., and Price, J. (2024). Teaching teachers to use computer assisted learning effectively: Experimental and quasi-experimental evidence. Working Paper 32388, National Bureau of Economic Research.
- Pape, B. (2018). Learner variability is the rule, not the exception. White paper, Digital Promise Global. Accessed: 2025-01-08.
- Rivkin, S. G., Hanushek, E. A., and Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2):417–458.
- Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, 94(2):247–252.

- Rodriguez-Segura, D. (2021). Edtech in developing countries: A review of the evidence. *The World Bank Research Observer*, 37(2):171–203.
- Rose, E. K., Schellenberg, J. T., and Shem-Tov, Y. (2022). The effects of teacher quality on adult criminal justice contact. Working Paper 30274, National Bureau of Economic Research.
- Vargas, J. F. and Rozo Villarraga, S. V. (2024). Right to education : Forced migration and child education outcomes. Policy Research Working Paper Series 10720, The World Bank.
- Wager, S. and Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113(523):1228–1242.
- World Bank (2018). World Development Report 2018: Learning to Realize Education’s Promise. Technical report, World Bank, Washington, DC.

A Intervention Details

As explained in Section 3.2, the intervention in Bogotá included two components. First, treated teachers received access to *ShaIA*, an Artificial Intelligence (AI) ecosystem that supports lesson planning. Second, treated teachers were offered support (using workshops, visits, and webinars) aimed at improving the effective use of *ShaIA*. This appendix provides additional details on *ShaIA* (Appendix Section A.1) and on the support provided by *Mentú* and the *Secretaría de Educación del Distrito* (SED) (Appendix Section A.2), the local education institution in the city of Bogotá.

A.1 ShaIA

ShaIA is an AI ecosystem created by *Mentú* to strengthen the teaching and learning of mathematics in public schools in Bogotá. Its purpose is to support teachers in the planning, implementation, and reflection of lessons, promoting inclusive and evidence-based practices aligned with the priority learning goals established by the SED. *ShaIA* is not intended to replace teachers, but rather to complement them by providing evidence-based teaching strategies that foster effective mathematics learning while addressing learner variability. It does so by channeling language models through an explicit pedagogical architecture designed to promote two core components: learner variability and mathematical mindsets.

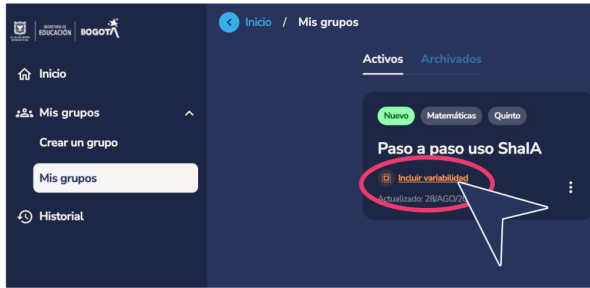
A.1.1 Learner variability

ShaIA's pedagogical approach is based on the premise that there is no such thing as an average student. Each classroom group combines diverse personal, cognitive, socio-emotional, and contextual factors that influence learning. Through a partnership with Digital Promise and its *Learner Variability Project*, *Mentú* designed *ShaIA* as a learning language model that translates this pedagogical vision into a practical process within a digital platform.

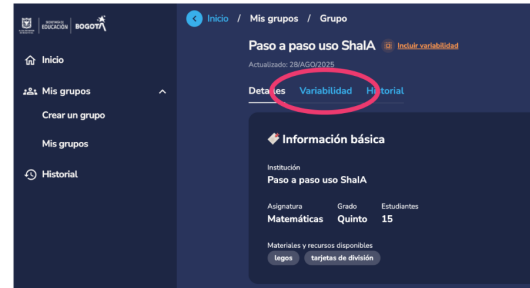
ShaIA incorporates learner variability for each of a teacher's courses. Individual access to the platform allows teachers to create a profile for each course by providing *ShaIA* with course-specific details. Appendix Figure A.1 presents a snapshot of the platform interface in which teachers are required to include learner variability in the algorithm.

Appendix Figure A.1: Course's Variability Profile Creation

Opción desde "mis grupos"



Opción desde información del grupo



For each course variability profile, teachers configure the inclusion of variability in three steps, through which inclusion ceases to be an abstract principle and becomes an active pedagogical configuration guiding every teacher interaction with the tool:

1. **Selection of the learning model and variability profile.** After the user selects the variability option, *ShalA* automatically suggests a learning model (e.g., Mathematics, Language, Adult Learning, or 21st Century Student) based on the group's information, which the teacher can confirm or adjust. Appendix Figure A.2 displays a snapshot of the platform where *ShalA* suggests a learning model but allows the teacher to modify it.

Appendix Figure A.2: Selection of Learning Model



Once the learning model is specified, teachers complete a questionnaire that col-

lects common variability aspects of the course.¹ This questionnaire collects information on students' Spanish language proficiency, migration background, exposure to traumatic events, and disabilities.² Appendix Figure A.3 displays a snapshot of the questionnaire, which is easy to complete using pre-determined options and an open-text field for additional information.

Appendix Figure A.3: Definition of Variability Profile

2. **Defining variability factors.** *ShalA* then uses this information to present a list of factors, which correspond to characteristics that can positively or negatively influence learning. These factors were co-developed by *Mentú* and *Digital Promise*, adapting the research on learner variability to the Bogotá context. Accordingly, *ShalA* presents a list of more than 35 factors identified by Digital Promise that impact learning, grouped into four domains:

- Personal background: physical well-being, sleep, socioeconomic status, adverse experiences.
- Socio-emotional skills: motivation, self-regulation, sense of belonging, emotion management.

¹The variability aspects included in the questionnaire were designed according to common indicators identified from the “Sistema Integrado de Matrícula” (SIMAT), which corresponds to the administrative student records in all Colombian schools.

²The tool also allows teachers to include open-text entries detailing additional conditions to take into account.

- Cognition: working memory, attention, metacognition, cognitive flexibility, visual processing.
- Mathematics: number sense, algebraic thinking, proportional reasoning, mathematical communication, among others.

Based on their knowledge of the group, teachers select between 10 and 15 factors that they consider most relevant to their students.

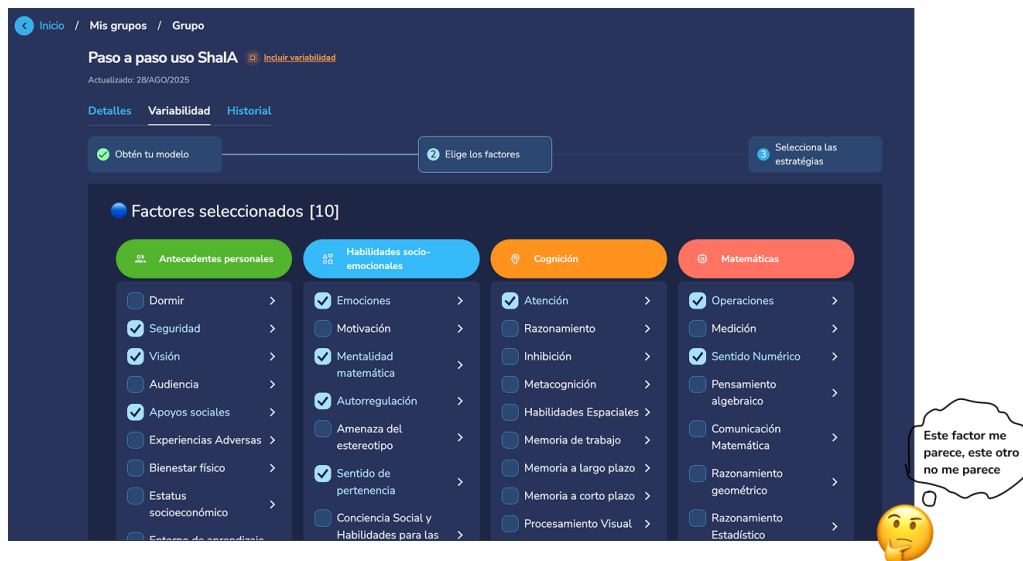
These factors go beyond traditional inclusion (i.e., disabilities in the traditional sense) and recognize that all students have strengths and challenges that vary depending on context. Variability factors cover a wide range and differ by grade or learning level (such as number sense and algebraic thinking). Specifically, for mathematics students in grades 6 through 11 (the target population of the intervention), *Digital Promise* has identified more than 90 classroom strategies that research has shown to support or strengthen 35 critical factors for learning. Table A.1 presents the learning factors adjusted to the intervention context:

Appendix Table A.1: Factors in the Learner Variability Model

Personal background	Socio-emotional skills	Cognition	Mathematics
Social support	Stereotype threat	Metacognition	Statistical reasoning
Physical well-being	Sense of belonging	Attention	Geometric reasoning
Hearing	Social and relational skills	Working memory	Algebraic thinking
Safety	Emotion	Long-term memory	Mathematical communication
Adverse experiences	Self-regulation	Inhibition	Proportional reasoning
Socio-economic status	Motivation	Visual processing	Measurement
Vision	Mathematical mindset	Processing speed	Operations
Sleep		Cognitive flexibility	Mathematical flexibility
Mathematics learning environment		Short-term memory	
		Reasoning	
		Spatial reasoning	

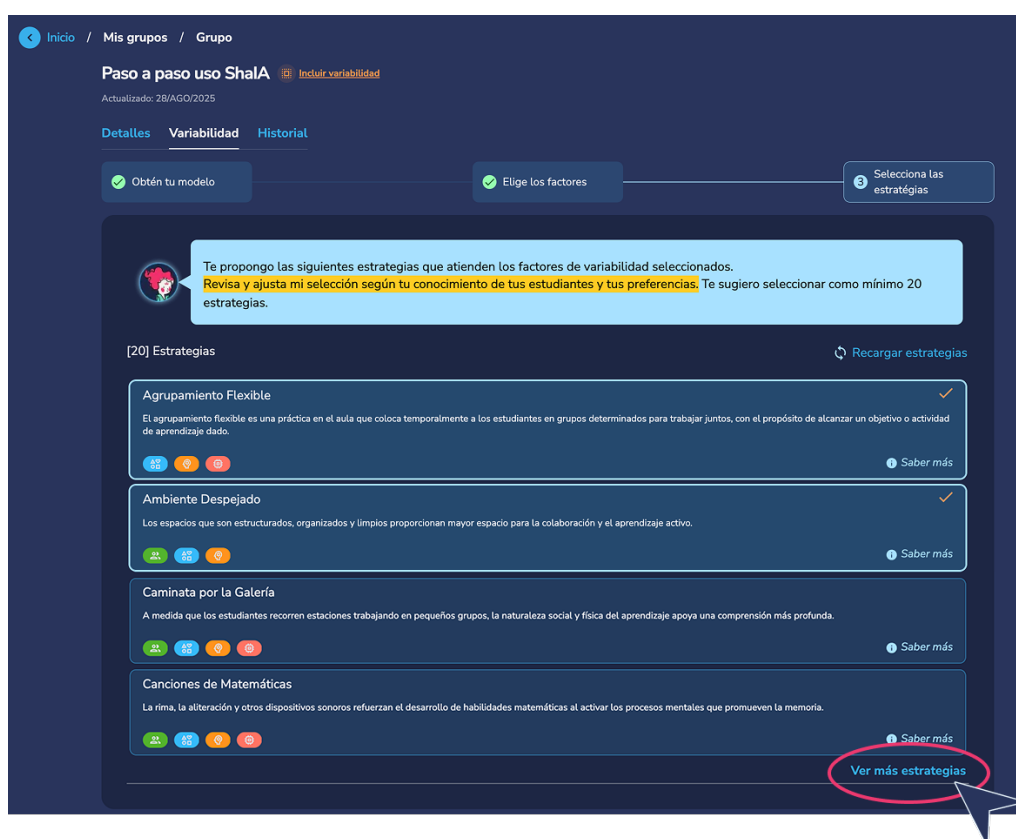
Appendix Figure A.4 presents *ShaIA*'s interface in which teachers select the factors they consider most relevant for their course. *ShaIA* provides an automatic selection of factors based on the information the teacher entered, but teachers can modify these selections based on their own professional judgement. The tool recommends prioritizing between 10 and 15 factors.

Appendix Figure A.4: Selection of Prioritized Factors



3. **Selection of inclusion strategies.** For each factor, *ShaIA* then suggests evidence-based pedagogical strategies (also curated by *Digital Promise* and adapted by *Mentu*) that allow for the simultaneous addressing of classroom diversity. Each of these strategies is a pedagogical practice that supports or enhances one or more factors by adapting instruction to students' actual needs. *ShaIA* again provides a list of strategies related to the previously selected factors. Teachers can modify this list and select the strategies they consider appropriate to implement. The tool recommends selecting at least 20 strategies so that *ShaIA* has a sufficiently broad repertoire to draw from. Appendix Figure A.5 presents a snapshot of selected teaching strategies for a specific course profile.

Appendix Figure A.5: Selection of Teaching Strategies



The curated evidence from Digital Promise offers a wide range of 95 teaching strategies adjusted for mathematics learning during lower secondary education (i.e., grades 6–9).³ These strategies—for example, making students’ ideas visible, using multiple representations, anticipating diverse strategies, or encouraging collaboration—are automatically incorporated by *ShaIA* into the generation of plans, activities, and resources, ensuring that the AI’s responses are aligned with the actual profile of the group.

A.1.2 Math Mindsets

ShaIA’s suggested teaching strategies are complemented by a set of mathematics-specific instructional tools that support teachers in the design and improvement of their lessons. These tools are aligned with the SED’s prioritized curriculum and with the Mathematical Mindsets approach developed by *YouCubed* (Boaler and Dweck, 2016). This approach proposes that all learners can achieve high levels of mathematical understanding through open, collaborative, and meaningful challenges. It consists of a library of educational resources aimed at teachers, with the goal of promoting a mathematical mindset focused on big ideas, complex thinking, and reducing mathematics anxiety. As part of the partnership between *Mentú* and *YouCubed*, these resources were translated into Spanish for the first time and made available to teachers participating in the intervention.

³A full list of the teaching strategies and their connection to learning factors can be found online at: <https://lvp.digitalpromiseglobal.org/content-area/math-7-10/str>.

The ecosystem for developing mathematical skills consists of five specific pedagogical tools: pattern conversation, numerical conversation, math experience, worksheet, and feedback practice. These five tools focus in particular on mathematical communication, reasoning, problem solving, and modeling, while also contributing indirectly to the development of procedural fluency. For each of these tools, *ShaIA* incorporates a set of teaching practices (including variability strategies from the learner variability component) that are aligned with the curriculum of the city of Bogotá. Appendix Table A.2 presents each pedagogical tool, the associated teaching practices, their relationship to the curriculum, and specific examples.

ShaIA integrates the Mathematical Mindsets approach to reduce mathematics anxiety, strengthen student confidence, and promote a connected view of mathematics as a system of ideas rather than a set of isolated rules. This approach is coupled with the learner variability component, which is pre-specified by the teacher.

Appendix Table A.2: Mathematical Mindsets Tools and Teaching Practices

Tool	Promoted Teaching Practice(s)	Relationship with Curriculum Prioritization	Example of product generated by <i>Shaia</i>
Pattern conversation	<ul style="list-style-type: none"> • Reducing math anxiety • Class agreements that promote participation and a growth mindset • Movements and conversations that reinforce students' ideas • Ability to paraphrase students' thinking • Strategies for building on students' ideas • Strategies for inviting contrast between ideas • Strategies for including variability • Anticipating students' strategies 	<p>Prioritised learning: Learning focused on generalisation and algebraic thinking, typically linked to the metric–spatial component.</p> <p>Mathematical thinking skills: Enhances reasoning and communication, as well as the ability to model situations. Indirectly contributes to procedural practice through the exploration of algebraic relationships.</p> <p>Components: Focuses on the metric–spatial component by working with visual representations, figure growth, and geometric patterns.</p>	Example 1
Numerical conversation	<ul style="list-style-type: none"> • Developing students' numerical flexibility • Reducing math anxiety • Class agreements that promote participation and a growth mindset • Movements and conversations that reinforce students' ideas • Ability to paraphrase students' thinking • Strategies for building on students' ideas • Strategies for inviting contrast between ideas • Strategies for including variability • Anticipating students' strategies 	<p>Prioritised learning: Contributes to learning linked to the numerical–variational component, strengthening numerical sense and relationships between operations.</p> <p>Mathematical thinking skills: Develops communication, reasoning, and procedural practice through numerical flexibility.</p> <p>Components: Focuses primarily on the numerical–variational component.</p>	Example 1
Mathematical experience	<ul style="list-style-type: none"> • Build scaffolding for open-ended tasks • Select different types of maths activities (visualise, solve, discuss, review, connect, play, explore) • Promote collaborative work • Strategies for including variability 	<p>Prioritised learning: Allows explicit selection of prioritised learning and evidence as the central objective.</p> <p>Mathematical thinking skills: Integrates communication, reasoning, problem solving, and modelling.</p> <p>Components: Can address numerical–variational, metric–spatial and random components in an integrated or isolated manner.</p>	Example 1
Worksheet	<ul style="list-style-type: none"> • Invite students to create graphic representations • Encourage collaborative work • Connect mathematical procedures and big ideas • Strategies for including variability 	<p>Prioritised learning: Allows you to set objectives and evidence directly associated with the SED's prioritised learning.</p> <p>Mathematical thinking skills: Promotes reasoning, communication, modelling, and problem solving.</p> <p>Components: The worksheet adapts to the prioritised learning selected by the teacher.</p>	Example 1 Example 2 Example 3
Feedback practice	<ul style="list-style-type: none"> • Feedback and growth mindset messages • Error management • Recognition of variability factor • Strategies for including variability • Making students' thinking visible 	<p>Prioritised learning: Enables teachers to offer formative feedback connected to prioritised learning.</p> <p>Mathematical thinking skills: Reinforces mathematical communication and reasoning.</p> <p>Components: Adapts to the learning component worked on in class.</p>	Example 1

A.2 Additional Support

The intervention was complemented by pedagogical support on the use of *ShaIA* provided by *Mentu*. This support was delivered through an in-person kick-off boot camp session, four on-site visits, three practical workshops, and three webinars. The kick-off session took place on August 14, 2025, at the beginning of the intervention. The visits and workshops were distributed throughout the year, while the webinars were scheduled during the first two months of the intervention.

1. **Kick-off Session.** This session consisted of a pedagogical and technological boot camp. It was a six-hour in-person working session that brought together teachers from participating schools in a shared space. The meeting combined plenary sessions with group work activities. Approximately 64% of the treated institutions attended the session. The day was divided into six sequential stages:
 - (a) The framework was presented, and the mathematics goals and prioritized learning outcomes were outlined. *ShaIA* was compared with other AI tools such as ChatGPT.
 - (b) Mathematical myths were addressed through a participatory activity, and the principles of the Mathematical Mindsets approach were presented.
 - (c) The concept of learner variability was introduced. Teachers analysed student profiles and proposed strategies based on the variability navigator. They then logged into *ShaIA* to create and parameterize a group based on these variability factors.
 - (d) A classroom practice was modelled through a numerical conversation and a low-floor, high-ceiling activity. Teachers participated from the role of “student” and then reflected from the role of “teacher.”
 - (e) A live demonstration of *ShaIA*’s mathematics tools (types of conversations, experiences, worksheets, and formative feedback) was conducted, showing how these facilitated the implementation of the Mathematical Mindsets approach.
 - (f) Individual commitments were formulated and shared regarding the implementation of innovative AI-based practices. The session concluded with a presentation of the next steps in the support programme and the administration of a closing survey.
2. **On-site visits and workshops.** The visits and workshops were delivered in an interconnected sequence, following the structure:

Visit 1 → Visit 2 → Workshop 1 → Visit 3 → Workshop 2 → Workshop 3 → Visit 4.

Schools that did not participate in the kick-off session received the same information during visit 1. In practice, similar activities were carried out in both visits and workshops, including classroom modelling (observed by participating teachers), group analysis of student work and evidence, presentations on the project’s pedagogical framework, and strategies for designing “low-floor, high-ceiling” activities. However, workshops were designed to train teachers

in modelling strategies and activities using *ShaIA*, whereas visits were intended to provide feedback based on classroom observations and to review teachers' portfolios and work plans.

More specifically, each component included:

- (a) **Visit 1.** The purpose of this visit was to introduce *ShaIA* and establish the institutional and pedagogical framework of the process. *ShaIA* was presented, the intervention was contextualized relative to the SED's mathematics goals, and existing practices, institutional conditions, and initial needs were identified.
- (b) **Visit 2.** This visit aimed to strengthen the appropriation of the pedagogical approach and translate it into a concrete implementation plan for the duration of the project. Teachers identified practices to implement and defined individual work plans involving commitments to apply them in the classroom. They also identified pedagogical and logistical barriers and discussed strategies for overcoming them.
- (c) **Visit 3.** The objective was to implement an activity designed with *ShaIA* in the classroom (with numerical conversations suggested) and to reflect on the practice through observation and co-observation exercises, followed by peer analysis sessions using a classroom observation protocol. The protocol examined practices such as precision in mathematical language, the number of students engaged in mathematical discussion, invitations to participate, reflection time, and peer communication.
- (d) **Visit 4.** This session combined a group reflection space with individual feedback sessions based on the consolidation of evidence portfolios from the process. The objective was to identify strategies for sustaining the work. Feedback on the programme offering was also collected.
- (e) **Workshop 1.** This workshop modelled a numerical conversation as an example of a low-floor, high-ceiling activity. Teachers participated as students and then used a protocol to analyse pedagogical aspects of the strategy (error management and student participation, attentive listening, communication of growth-mindset messages, and visibility of student thinking). Teachers then used *ShaIA*'s Numerical Conversations tool to plan a numerical conversation to implement with their students and received feedback from the facilitator during the subsequent visit.
- (f) **Workshop 2.** The objective was for teachers to analyse student work using a protocol to identify evidence of processes such as thinking strategies, errors, opportunities for agency, and spatial reasoning. These findings were intended to provide criteria for adjusting classroom activities and to support understanding of the value of *ShaIA*'s Mathematical Experience and Worksheet tools.
- (g) **Workshop 3.** The objective was for teachers to transform closed tasks (single answer, single point of entry, procedural application) into open activities using five specific strategies presented in the workshop. In addition to sharing strategies for opening tasks, the workshop aimed to help teachers identify the characteristics of low-floor, high-ceiling activities, thereby strengthening understanding of the pedagogical intent of *ShaIA*'s mathematics tools.

Although these visits and workshops could not be scheduled simultaneously across all schools, it was expected that by the end of the intervention all treated schools would receive the full support package, including four on-site visits and three workshops.

3. **Webinars.** This component sought to create learning spaces that fostered a sense of belonging and motivation toward the programme through one-hour hybrid sessions offered after school hours. Three webinars were offered:
 - (a) **Webinar 1.** Held on September 4, 2025, its objective was to introduce and deepen understanding of *ShaIA*'s pedagogical offering, including its tools, factors, and strategies. Emphasis was placed on aligning tools (such as Mathematical Conversations and Open Worksheets) with the skills prioritized by the SED.
 - (b) **Webinar 2.** Held on September 25, 2025, this session addressed the “what, how, why, and for what” of prioritizing learning in mathematics. Challenges identified in external assessments (Saber 11, Saber 3rd, 5th, and 9th) revealing gaps in argumentation and representation were contextualized. Participants practiced by designing a worksheet in *ShaIA*, using prioritized learning objectives as a reference.
 - (c) **Webinar 3.** Held on October 23, 2025, this session sought to strengthen teachers' capacity to integrate AI in a critical and ethically sound manner. Risks such as bias, hallucinations, and cognitive debt were analysed. The importance of “best prompts” was presented to ensure that *ShaIA*'s outputs are pedagogically sound, linking effective prompting with prioritized learning and variability strategies.