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

Preliminary Results

Felipe Barrera-Osorio and Juan Muñoz-Morales *

June 19, 2026

Click here for the latest version

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. Using baseline and midline student assessments complemented with teacher surveys, we report preliminary midline results based on a subsample of schools. The estimates suggest that the intervention improves math achievement by around 0.2 to 0.4 standard deviations. These results are cost-effective relative to prior evidence on educational interventions. They should be interpreted as exploratory pending endline data collection.

Keywords: Artificial Intelligence, Classrooms, Math, Treatment Effect Heterogeneity, Colombia.
JEL classifications: C93, I21, J16, J24, O15, O33.

Study pre-registration: AEARCTR-0017147

*Barrera-Osorio: Vanderbilt University (felipe.barrera.-.osorio@vanderbilt.edu). Muñoz-Morales: Univ. Lille, CNRS, IESEG School of Management, UMR 9221 - LEM - Lille Économie Management, Lille F-59000, 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 socioeconomic background, prior skills (including cognitive and socio-emotional), and learning pace, among others. As a result, interventions that raise average classroom performance may leave the lowest-performing students behind (Buhl-Wiggers et al., 2024). 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 an AI–enabled, teacher-level intervention designed to address within-classroom heterogeneity in student learning. We partner with *MentuLabs*, 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. We designed 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 and expanding the repertoire of teaching strategies available to teachers, both key constraints 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, where within-classroom learning gaps are large and institutional support for differentiated instruction is limited.

The intervention is implemented in Bogotá, Colombia, 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.

The evaluation was conducted in 158 eligible public schools located in low-income areas of Bogotá, which predominantly serve disadvantaged students with limited access to high-quality educational resources. The intervention was rolled out through a school-level randomized design. Of the 158 schools, 70 were originally assigned to treatment, where all mathematics teachers in grades 6 through 9 were offered access to

[1]The 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.

ShaIA together with complementary training designed to promote effective use of the platform.

Participation was voluntary for both schools and teachers. To improve balance and precision, we grouped schools into 52 blocks based on pre-intervention characteristics and randomly assigned treatment within each block. In some blocks a school assigned to treatment did not participate. In a few of these, a school originally assigned to control was reassigned to treatment, breaking random assignment within the block but preserving the randomization in other blocks. We drop schools in blocks where a control school became treated, and work with two samples. The larger sample retains the remaining blocks, including those where a treated school did not participate, and the pure sample keeps only the blocks where the original assignment was preserved, leaving 40 treated and 50 control schools. Because teacher take-up within treated schools was also voluntary, our empirical strategy estimates both intention-to-treat and treatment-on-the-treated effects.

Student learning outcomes were measured at baseline in August 2025 and at midline in November 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.

The midline collection covers a non-random subsample of 29 schools (15 treated and 14 control), which we use to provide a first look at the intervention's effects on student test scores. These schools were drawn from across the full sample of 158 schools, not from the pure-sample blocks where randomization was preserved and not in a way that maintains within-block coverage. This sampling raises two concerns about the validity of the estimates, one from the subsample not being representative and one from the loss of the within-block structure. The first concern is unfounded, as the sampled schools are similar to the rest of the sample on a wide range of observable characteristics, which supports a cross-sectional comparison of treated and control schools. We adopt this comparison as our preferred specification at midline. The second concern remains, since within-block comparability is weakened when only a few blocks contain both a treated and a control school observed at midline, so we report the within-block estimates as a robustness check rather than as our main result.

Treatment effects on math test scores are positive in our preferred cross-sectional comparison, on the order of 0.2 standard deviations. Within-block comparisons yield higher point estimates, on the order of 0.4 standard deviations, but rest on a small number of schools and blocks where the random assignment property is compromised, and should therefore be interpreted with caution. The effect is concentrated in the probability and statistics component and the problem-solving competence of the test rather than spread evenly across content areas. Treatment-on-the-treated estimates

that instrument program take-up with assignment point in the same positive direction, although less precise because of a loss in sample size. These magnitudes are in line with effect sizes documented in recent randomized evaluations of AI-based educational interventions. Because they rest on a non-random subsample and on a small number of schools, these midline estimates should be interpreted as preliminary and will be revisited at endline.

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; Muralidharan 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

[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.

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:

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

[3]For 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.

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 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.

Building on this framework, we test the role of AI in improving 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 channels. First, AI can reduce the cost of class preparation and pedagogy delivery, which in turn can 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. Because AI can enhance teachers' capacity to tailor instruction to diverse learning profiles, we expect the intervention to generate larger treatment effects for historically disadvantaged groups and thereby reduce within-classroom variability in learning outcomes. In the setting we analyze, girls and migrant students exhibit significantly lower baseline performance in math, and we anticipate that the intervention will allow teachers to better support these groups while raising overall learning.

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, girls, 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 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 girls 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 the 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 partnership with the *Secretaría de Educación Distrital* (SED)—the local education authority in Bogotá,

[4]Since 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).

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 diversity simultaneously.[6] These strategies are automatically incorporated into

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

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

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, *Mentu* and the SED implemented a structured training program for treated teachers. The program consisted of an in-person kick-off bootcamp session, three practical workshops, four on-site visits, and three webinars. The training was designed not only to familiarize teachers with the platform, but also to support a broader shift in instructional practice toward the differentiated pedagogy embedded in the learner variability framework. To this end, the workshops modeled complete lesson plans designed with *ShaIA* using this approach, illustrating how the platform can be used to structure instruction around heterogeneous student needs. The visits were intended to gather feedback and to review the teachers' work plan. Finally, the

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

webinars were designed to reinforce *ShaIA*'s instructional applications and support its integration into classroom practice.[8]

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 (5) of panel A), 58 schools were offered treatment from a pool of 98 eligible schools grouped into 40 blocks. In this

[8] Appendix A.2 provides more details on the available resources offered to treated schools.

[9] All 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.

round, blocks consisted of two or three schools where one control school was assigned. In the second round (described in columns (6) to (10) 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.

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 (four of which are out of the evaluation), but randomization was compromised in 14 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. This variation is used to explore heterogeneity in treatment intensity and timing.

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.

4.2 Data Collection

Figure 1 outlines the implementation of the evaluation, which took place during 2025 and 2026. Randomization and treatment assignment were conducted during August 2025. Treated schools received access to ShaIA beginning in the second semester of 2025 and continuing through 2026. The evaluation concludes at the end of the academic year, before November 2026.

Students were evaluated and teachers were surveyed at three 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). A midline measurement was collected in November 2025, in a smaller *non-random* subsample; we describe its composition and limitations in Section 4.3. 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 surveys include sociodemographic characteristics, information on teaching practices, their use of digital technologies, and their perceptions of AI. Importantly, at baseline, teachers were asked whether they knew about ShaIA, but no additional information about the platform was collected, as it is expected that neither teachers in treated nor control schools were 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.

Within each school, classrooms are randomly selected for testing at every wave of data collection, so not all students are observed. For the students in these classrooms, we collect 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.

[10] The 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 September or November 2026.

[11] We 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 project.

All test items were pre-tested in an alternative setting to ensure their validity and suitability for the local student population. Student data was 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]

4.3 Midline Data Collection

A midline measurement was collected at the end of 2025 in a subsample of 29 schools (15 treated and 14 control), covering 2,432 students. This collection was implemented under tight operational constraints. Data had to be gathered during a short window before the end of the academic year, and scheduling difficulties limited the number of schools that could be reached. As a result, the selection of schools into the midline subsample was not random. It was not restricted to the pure sample, nor designed to ensure that within each randomization block both a treated and a control school were observed. The midline schools were instead drawn from the universe of eligible schools, including schools belonging to blocks where randomization was compromised by non-participants or switchers. We additionally collected 72 voluntary teacher surveys, of which 28 could be merged to their students.

Table 3 summarizes the structure of the midline sample. Panel A reports the total number of midline schools under three sample restrictions, namely the universe of midline schools, the sample excluding control schools that were reassigned to treatment, and the pure sample. Panel B restricts these three samples to schools belonging to blocks where at least one treated and at least one control school are jointly observed at midline.

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:

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

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

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

and zero otherwise, and γ_b represents block fixed effects. Standard errors are clustered at the school level.

Some teachers in treated schools may choose not to participate in the intervention. As a result, some units assigned to treatment may remain untreated, implying that Equation 1 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:

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

Midline Estimation Strategy. The non-random selection of schools into the midline subsample raises two distinct concerns about the validity of the estimation strategy described in Equation 1. The first is that schools could have been selected on characteristics correlated with potential outcomes. The second is that the within-block structure of the original randomization could be lost. Our evidence indicates that the first concern is largely unfounded, but that the second is binding. Appendix Table A.1 compares the 29 midline schools to the rest of the sample not included at midline on a wide set of school-level administrative characteristics. The two groups are statistically similar on most observables, with only a few isolated differences that are unlikely to reflect systematic selection. The midline schools therefore appear to be a reasonably representative subset of the experiment along observable dimensions.

By contrast, the within-block structure of the midline sample is incomplete. As shown in Panel B of Table 3, only 8 of the 29 midline schools belong to blocks where both a treated and a control school are observed at midline, spanning 4 blocks. Of these 4 blocks, only one is pure, consisting of one treated and one control school where randomization was implemented as designed. The remaining three blocks are non-pure, meaning that both a treated and a control school are observed but at least one treated school within the block declined participation. Within-block comparisons therefore remain feasible only in these 8 schools. However, except for the single pure block, the within-block contrast no longer compares schools drawn at random from the same block, which is the property that justified the design.

Cross-sectional comparisons across all midline schools, restricted to blocks where control schools did not become treated, are therefore the preferred midline specification, provided that baseline balance holds across these schools.[13] Within-block comparisons

[13] We assess this in Section 5.1.

at midline are therefore best read as a robustness check. Student- and school-level controls are included in the conditional specifications to mitigate potential residual imbalances at midline.

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 (Appendix Table A.2) and baseline survey data (Appendix Table A.3). 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. The balance is robust along two complementary dimensions. The unadjusted p-values are insignificant for the vast majority of variables, indicating that treated and control schools are comparable on average even before exploiting the block structure of the design. The block-adjusted p-values are also largely insignificant, confirming that the within-block contrast that motivated the randomization is preserved in the data.

The strong unconditional balance is particularly relevant for the midline analysis described in Section 5, where the limited within-block coverage makes the cross-sectional contrast the natural identification strategy. Appendix Tables A.4 and A.5 present the school-level administrative balance and the student- and teacher-level balance restricted to the 29 midline schools. The two groups are balanced on baseline characteristics in unconditional comparisons, but a number of variables become imbalanced once randomization-block fixed effects are introduced. This pattern is a direct consequence of the incomplete within-block coverage of the midline sample: many blocks contribute only one treated or one control school at midline, so block fixed effects absorb most of the identifying variation, leaving the residual contrast unstable.

6 Midline Results

We begin by documenting program take-up before turning to the student outcome estimates. Table 4 confirms that control school teachers had no access to ShaIA, while among treated schools around 76 percent used the platform at least once and the

overall training attendance rate was approximately 51 percent.[14] Fewer than one third completed the minimum attendance threshold of 70 percent, reflecting the partial compliance that motivates the treatment-on-the-treated estimates below.

Table 5 reports ITT estimates of β_{ITT} from Equation 1 on the overall test score, organized in two panels. Panel A uses baseline test scores for the schools later observed at midline as a balance check, and Panel B reports midline test scores. Columns (1) to (3) cover the sample excluding schools in blocks where control schools were reassigned to treatment, and columns (4) to (6) the pure sample. Within each group, the first column reports estimates without controls, the second adds student-level controls, and the third includes block fixed effects as a robustness check.[15]

Two patterns emerge. First, treated and control schools observed at midline are balanced on baseline test scores in the unconditional comparisons of Panel A (columns (1), (2), (4), and (5)), but a clear imbalance appears once block fixed effects are introduced. As discussed in Section 5, this pattern arises because incomplete within-block coverage leaves the residual contrast unstable once block-level variation is absorbed by the fixed effects.

Second, the unconditional and control-adjusted estimates in Panel B point to positive treatment effects on midline test scores. In the sample excluding switchers, the estimate is around 0.20 standard deviations and is statistically distinguishable from zero. In the pure sample, the estimates fall in magnitude but remain positive, although precision is lost. A useful way to read these numbers is in terms of the change in the treated-control gap between baseline and midline. In both samples, this gap moves from a small negative value at baseline to a positive value at midline. The implied shift is remarkably stable across the four columns, at 0.22, 0.24, 0.17, and 0.24 standard deviations for columns (1), (2), (4), and (5) without block fixed effects, which we read as evidence of a consistent positive effect of the intervention on math achievement.

The block-fixed-effects estimates in columns (3) and (6) of Panel B are substantially larger. Read as the change in the treated-control gap between baseline and midline, they yield shifts of 0.41 and 0.49 standard deviations across the two samples, well above the 0.17 to 0.24 standard deviations implied by the specifications without block fixed effects. We interpret these magnitudes as upper bounds rather than central estimates, since adding block fixed effects inflates both the baseline imbalance in Panel A and the midline gap in Panel B in the same direction, consistent with the unstable within-block contrast discussed in Section 5.[16]

[14] We do not have administrative data on the universe of math teachers in each school. The sample is therefore defined as those identified through at least one data source: the teacher survey, platform records, or attendance logs from workshops, seminars, and bootcamps.

[15] Student-level controls include a student-level wealth index, gender, and migrant status.

[16] The block-fixed-effects estimate in column (6) is therefore identified almost entirely from a single pair comparison of schools, which makes the magnitude implausibly large relative to the range documented in the literature on comparable interventions.

Table 6 reports the same midline estimates separately for the three curricular components (probability and statistics, spatial-metric, and numerical-variational) and the three competences (communication, problem solving, and reasoning). The positive effect is concentrated rather than spread evenly across the test. Among components, it is driven by the probability and statistics component (0.245 standard deviations, bootstrap p-value of 0.020). Among competences, it loads on problem solving (0.272 standard deviations, bootstrap p-value of 0.036), while communication and reasoning are near zero. Panel A shows no corresponding baseline imbalance, supporting a treatment-effect interpretation.

Table 7 reports treatment-on-the-treated estimates, instrumenting two measures of program take-up: training attendance rate and number of platform uses.[17] The training attendance instrument is strong, and the corresponding estimates imply that a one-unit increase in the attendance rate raises midline test scores by around 0.4 standard deviations, positive but not statistically distinguishable from zero. The platform-uses estimates are positive and statistically significant but rest on a weak first stage (F between 2 and 4), so we read them with caution.[18] Taken together, the TOT estimates are consistent with the ITT results and suggest that the effects are driven by teachers who actively engaged with the program, though the limited sample size prevents precise inference at this stage.

The magnitudes of these midline estimates are consistent with effect sizes documented in recent randomized evaluations of AI-based educational interventions. De Simone et al. (2025) report effects of 0.2 to 0.3 standard deviations from a six-week student-level intervention using a generative AI assistant in Nigeria, and Henkel et al. (2024) report effects of 0.37 standard deviations from an eight-month AI-based tutor in Ghana. Oreopoulos et al. (2024) report effects of 0.12 to 0.22 standard deviations from a teacher-level computer-assisted learning program in the United States and Canada, and Andrabi et al. (2025) report effects of 0.12 standard deviations from an eight-week app-based teacher-support intervention in Pakistan.

These estimates should be interpreted as suggestive. Although the 29 midline schools appear similar to the rest of the experimental sample on observable characteristics, as documented in Appendix Table A.1, the within-block coverage at midline is incomplete (Section 4.3). Identification at midline therefore relies on across-block variation between treated and control schools, rather than on the strict within-block design originally intended. The results provide an early indication of the direction

[17] We merge teacher survey information with student test scores for 72 surveyed teachers. Because midline test scores were collected in randomly selected classrooms within each school, not all surveyed teachers can be matched to student outcomes. We successfully match 28 teachers, 12 treated and 16 control, who taught 847 students.

[18] We do not report estimates with block fixed effects in the TOT specification, as the within-block variation is insufficient to yield stable estimates given the limited number of schools observed at midline.

and order of magnitude of the intervention's effects, but do not yet constitute the full design-based estimates that will be available at endline.

7 Cost-Effectiveness and Conclusion

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.[19] On average, each math teacher in the sample teaches 120 students, resulting in a total implementation cost of $9.14 USD per student, of which $2.20 corresponds to licenses and $6.94 to support.

Combining this implementation cost with the midline effects of 0.2 to 0.4 standard deviations reported in Section 6 yields a cost of roughly $2.30 to $4.60 USD per 0.1 standard deviations of math achievement gained, which compares favorably to other interventions in the literature on educational technology and instructional quality. For reference, the average effect size across more than 700 randomized evaluations of education interventions is approximately 0.10 standard deviations (Kraft, 2020), so our midline estimates are two to four times the typical effect in this literature even before accounting for cost. Targeted-instruction programs in the spirit of Teaching at the Right Level have produced effects of 0.15 to 0.30 standard deviations in India and several African countries at a cost typically below $10 per student per year, ranking among the most cost-effective education programs evaluated to date (Banerjee et al., 2017; Duflo et al., 2024). The outsourcing of public schools to private providers in Liberia produced effects of 0.18 standard deviations at an incremental cost of $50 per pupil (Romero et al., 2020). Closer to our setting, EdTech interventions delivering instructional content via short video lessons in Pakistan have produced effects of around 0.30 standard deviations (Beg et al., 2022), and low-cost phone-based tutoring in Botswana produced effects of 0.12 standard deviations at a cost equivalent to roughly 0.9 standard deviations of learning per $100 spent (Angrist et al., 2022). Intensive in-person tutoring in the United States has been documented to deliver effects of 0.20 to 0.40 standard deviations, but at substantially higher per-student costs in the range of $1,000 to $2,500 (Nickow et al., 2020).

Relative to these benchmarks, our intervention combines effect sizes at the upper end of the distribution with implementation costs at the low end. This favorable cost-effectiveness profile reflects two features of the design. The intervention operates through teachers rather than students, leveraging the fact that each teacher serves many students at once, and it relies on a digital platform whose marginal cost of delivering content is low once the underlying infrastructure is in place. These results should

[19] The 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.

be interpreted as preliminary, as they rely on midline estimates from a non-random subsample of schools, and will be revisited at endline.

Figures and Tables

Figure 1: Timeline of Intervention and Data Collection

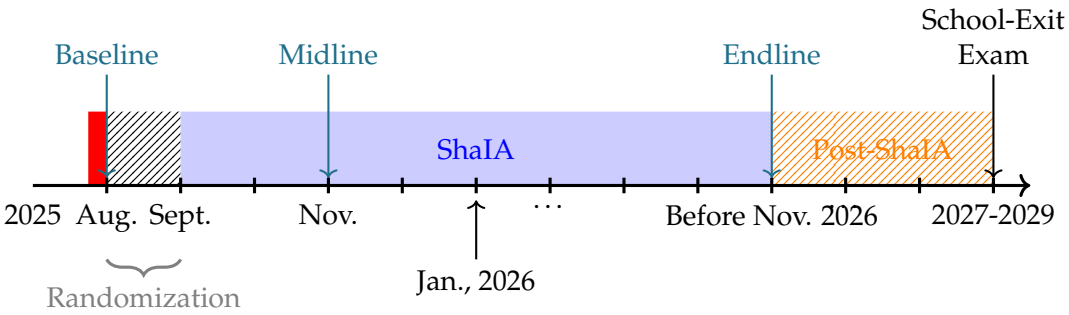


Table 1: Test Score Gaps in High School Exit Exam

	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		
	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)	Overall (13)
A) Original Randomization													
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
B) Implementation													
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: Midline Sample Composition

	Schools (1)	Treatment (2)	Control (3)	Blocks (4)	Students (5)
A) Total Collected					
Total	29	15	14	25	2,432
Dropping Switchers	28	14	14	24	2,324
Pure	20	11	9	19	1,644
B) T and C Observed within Block					
Dropping Switchers	8	4	4	4	710
Non-Pure	6	3	3	3	525
Pure	2	1	1	1	185

Note: The table reports the number of schools, treated schools, control schools, randomization blocks, and student observations in the midline sample. Panel A reports counts for all schools observed at midline, under three sample restrictions: the universe of midline schools (Total, n=29), the sample excluding control schools that were reassigned to treatment (Dropping Switchers, n=28), and the pure sample, which restricts to blocks where randomization was implemented as designed (Pure, n=20). Panel B restricts the same three samples to schools belonging to blocks where at least one treated and at least one control school are jointly observed at midline. Non-pure corresponds to schools in blocks where at least one treated school declined participation, but no control school was reassigned to treatment.

Table 4: Teacher-Level Platform Adoption and Training Attendance

	Dropping Switchers				Pure			
	Obs.		Mean		Obs.		Mean	
	Treated (1)	Control (2)	Treated (3)	Control (4)	Treated (5)	Control (6)	Treated (7)	Control (8)
Overall Attendance Rate (%)	153	99	0.51	0.00	132	69	0.51	0.00
Sessions and Workshops	153	99	0.54	0.00	132	69	0.53	0.00
Webinars	153	99	0.25	0.00	132	69	0.23	0.00
Completed 70% Attendance	153	99	0.30	0.00	132	69	0.28	0.00
Platform Use (%)	153	99	0.76	0.00	132	69	0.74	0.00
Total Platform Uses	116	0	9.09	.	98	0	8.49	.

Note: Each row reports the number of observations and mean of the indicated variable for treated and control schools. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table 5: Midline Treatment Effects on Overall Test Score

	Dropping Switchers			Pure		
	No FE	Controls	Block FE	No FE	Controls	Block FE
	(1)	(2)	(3)	(4)	(5)	(6)
A) Midline Schools in Baseline						
Treated	-0.020	-0.030	0.108**	-0.103	-0.107	0.296***
	(0.103)	(0.098)	(0.049)	(0.128)	(0.120)	(0.016)
Observations	2,354	2,354	2,354	1,676	1,676	1,676
Wild-Cluster BS	0.831	0.783	0.153	0.466	0.410	0.272
B) Midline Results						
Treated	0.198*	0.207**	0.521***	0.063	0.133	0.784***
	(0.115)	(0.098)	(0.053)	(0.142)	(0.119)	(0.015)
Observations	2,264	2,264	2,264	1,604	1,604	1,604
Wild-Cluster BS	0.105	0.056	0.120	0.689	0.311	0.331
Baseline Controls		Yes	Yes		Yes	Yes
Block FE			Yes			Yes

Note: The dependent variable is the overall standardized math test score. Panel A uses test scores at baseline, restricting the baseline sample to schools collected in midline. Panel B uses tests scores gathered in midline. Each column reports the coefficient on the treatment indicator from a regression of test scores on treatment assignment. Standard errors clustered at the school level are reported in parentheses. Columns labeled with *Controls* include student-level wealth, gender, and migrant status. The estimations in Panel B additionally includes school-level means baseline test scores. Columns labeled with *Block FE* include randomization block fixed effects. Columns (1) to (3) include the overall sample dropping schools in blocks were a control school was reassigned to treatment. Columns (4) to (6) restrict to blocks where randomization was implemented as designed. *** p<0.01, ** p<0.05, * p<0.1.

Table 6: Midline Treatment Effects by Component and Competence

	Components						Competences					
	Prob. and Stats.		Spatial-Metric		Numerical-Var.		Communication		Problem Solving		Reasoning	
	Drop.	Pure	Drop.	Pure	Drop.	Pure	Drop.	Pure	Drop.	Pure	Drop.	Pure
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
A) Midline Schools in Baseline												
Treated	0.013	-0.050	-0.036	-0.092	-0.044	-0.094	-0.066	-0.077	-0.032	-0.126	0.036	-0.024
	(0.085)	(0.107)	(0.088)	(0.123)	(0.073)	(0.076)	(0.075)	(0.101)	(0.089)	(0.100)	(0.090)	(0.113)
Observations	2,350	1,673	2,350	1,673	2,350	1,673	2,350	1,673	2,350	1,673	2,350	1,673
Wild-Cluster BS	0.894	0.643	0.673	0.493	0.568	0.241	0.423	0.556	0.713	0.239	0.695	0.818
B) Midline Results												
Treated	0.245**	0.192	0.056	0.016	0.110	0.053	0.085	0.061	0.272**	0.190	0.029	-0.013
	(0.097)	(0.124)	(0.069)	(0.075)	(0.072)	(0.085)	(0.082)	(0.108)	(0.105)	(0.118)	(0.062)	(0.080)
Observations	2,264	1,604	2,264	1,604	2,264	1,604	2,264	1,604	2,264	1,604	2,264	1,604
Wild-Cluster BS	0.020	0.215	0.447	0.858	0.154	0.586	0.344	0.617	0.036	0.145	0.658	0.880
Baseline Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Note: The dependent variable in each column is the standardized test score for the indicated component or competence. Panel A uses test scores at baseline, restricting the baseline sample to schools collected in midline. Panel B uses test scores gathered at midline. Each column reports the coefficient on the treatment indicator from a regression of test scores on treatment assignment, including student-level wealth, gender, migrant status, and school-level mean baseline test scores as controls. Standard errors clustered at the school level are reported in parentheses. Columns labeled *Drop.* use the sample dropping schools in blocks where a control school was reassigned to treatment. Columns labeled *Pure* restrict to the pure sample. *** p<0.01, ** p<0.05, * p<0.1.

Table 7: Treatment Effects on the Treated at Midline

	Dropping Switchers		Pure	
	No FE	Controls	No FE	Controls
	(1)	(2)	(3)	(4)
Training Attendance Rate	0.428	0.386	0.411	0.362
	(0.309)	(0.285)	(0.419)	(0.384)
Observations	847	847	502	502
Wild-Cluster BS	0.230	0.281	0.542	0.566
First-Stage F	81.522	90.103	35.122	38.721
Platform Uses (ShaIA)	0.011**	0.011**	0.008*	0.008
	(0.005)	(0.005)	(0.005)	(0.005)
Observations	847	847	502	502
Wild-Cluster BS	0.101	0.113	0.216	0.221
First-Stage F	2.713	3.030	1.998	2.173
Baseline Controls		Yes		Yes
Block FE				

Note: The dependent variable is the overall standardized math test score. Each column reports the 2SLS coefficient on the endogenous variable, instrumented by treatment assignment. Standard errors clustered at the school level are reported in parentheses. Wild-Cluster BS reports the wild cluster bootstrap p-value on the endogenous variable following **?**. First-Stage F reports the Kleibergen-Paap rk Wald F-statistic. Columns labeled with *Controls* include student-level wealth, gender, and migrant status, as well as school-level mean baseline test scores. Columns (1) and (2) use the sample dropping schools in blocks where a control school was reassigned to treatment. Columns (3) and (4) restrict to blocks where randomization was implemented as designed. Within each group, the first column reports estimates without controls and the second adds student-level and school-level controls.*** $p<0.01$, ** $p<0.05$, * $p<0.1$.

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.

Angrist, N., Bergman, P., and Matsheng, M. (2022). Experimental evidence on learning using low-tech when school is out. *Nature Human Behaviour*, 6(7):941–950.

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.

Beg, S. A., Halim, W., Lucas, A. M., and Saif, U. (2022). Engaging teachers with technology increased achievement, bypassing teachers did not. *American Economic Journal: Economic Policy*, 14(2):61–90.

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., Montaño, 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., Ibarrarán, 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, A., Kiessel, J., and Lucas, A. M. (2024). Experimental evidence on four policies to increase learning at scale. *Economic Journal*, 134(661):1985–2008.

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.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, 49(4):241–253.

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.

Nickow, A. J., Oreopoulos, P., and Quan, V. (2020). The impressive effects of tutoring on PreK-12 learning: A systematic review and meta-analysis of the experimental evidence. NBER Working Paper 27476, National Bureau of Economic Research.

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.

Romero, M., Sandefur, J., and Sandholtz, W. A. (2020). Outsourcing education: Experimental evidence from liberia. *American Economic Review*, 110(2):364–400.

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 *Mentu* 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 *Mentu* 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*, *Mentu* 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

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, *ShaIA* 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 *ShaIA* suggests a learning model but allows the teacher to modify it.

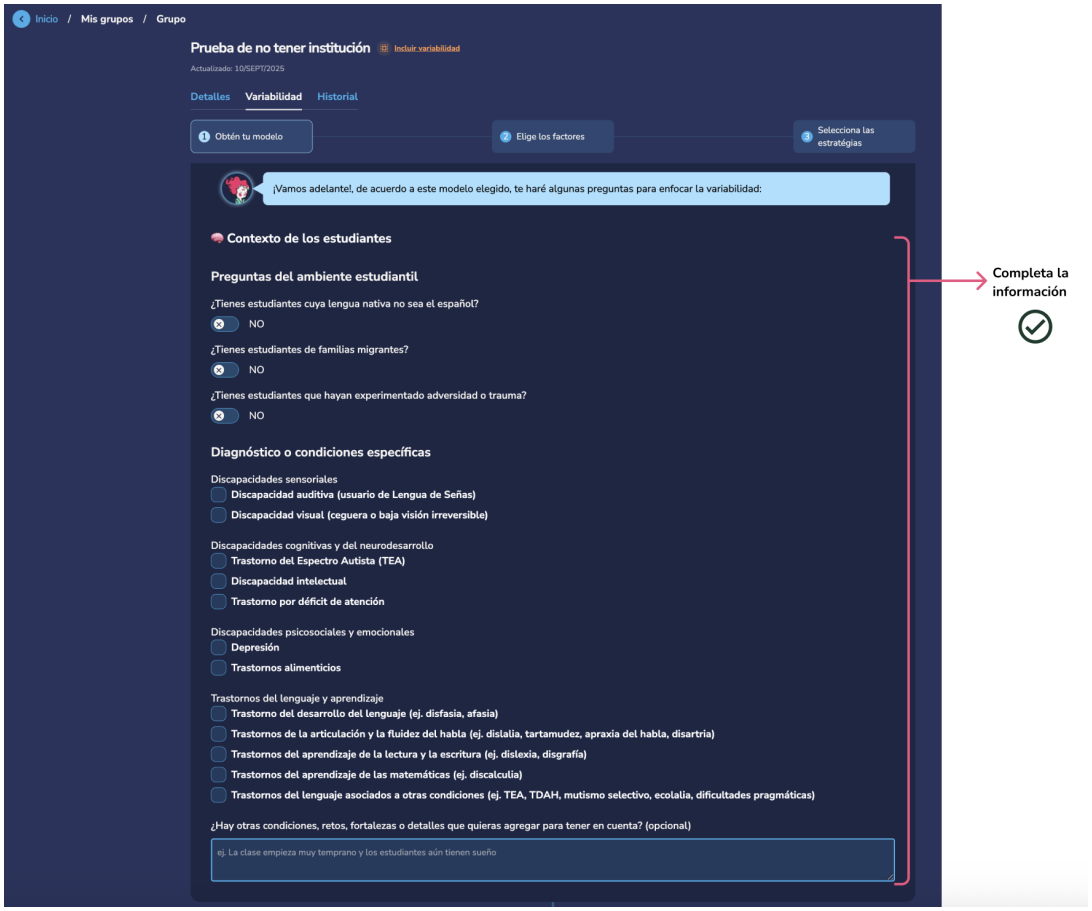
Appendix Figure A.2: Selection of Learning Model

Once the learning model is specified, teachers complete a questionnaire that collects common variability aspects of the course.[1] This questionnaire collects

[1]The variability aspects included in the questionnaire were designed according to common indicators

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.** *ShaIA* 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 *Mentu* and *Digital Promise*, adapting the research on learner variability to the Bogotá context. Accordingly, *ShaIA* 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.
    - Cognition: working memory, attention, metacognition, cognitive flexibility, visual processing.

identified from the "Sistema Integrado de Matrícula" (SIMAT), which corresponds to the administrative student records in all Colombian schools.

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

- 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.

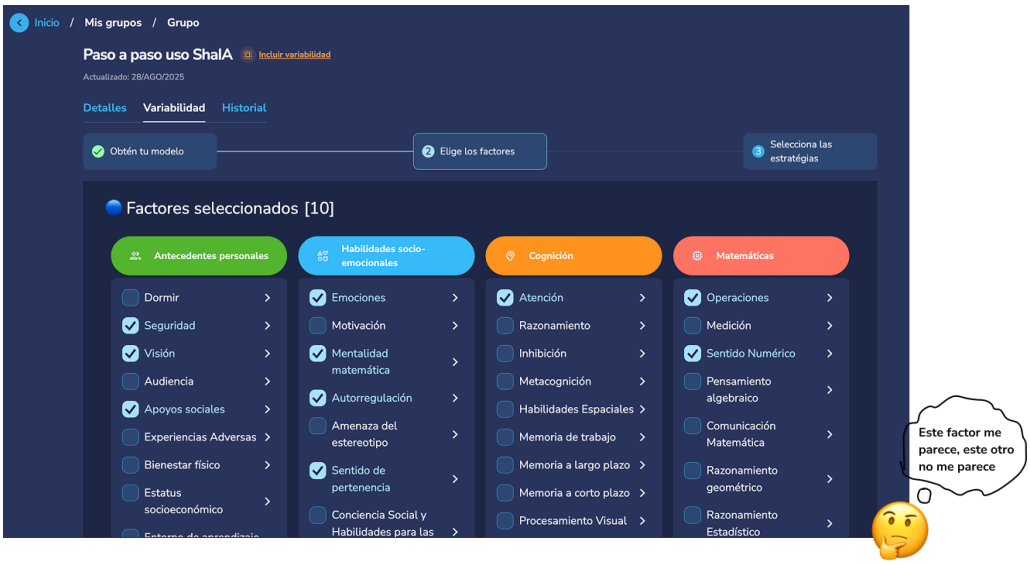
These factors go beyond traditional inclusion (i.e., disabilities in the traditional sense) and recognize that all students have strengths and challenges that vary depending on context. Variability factors cover a wide range and differ by grade or learning level (such as number sense and algebraic thinking). 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	

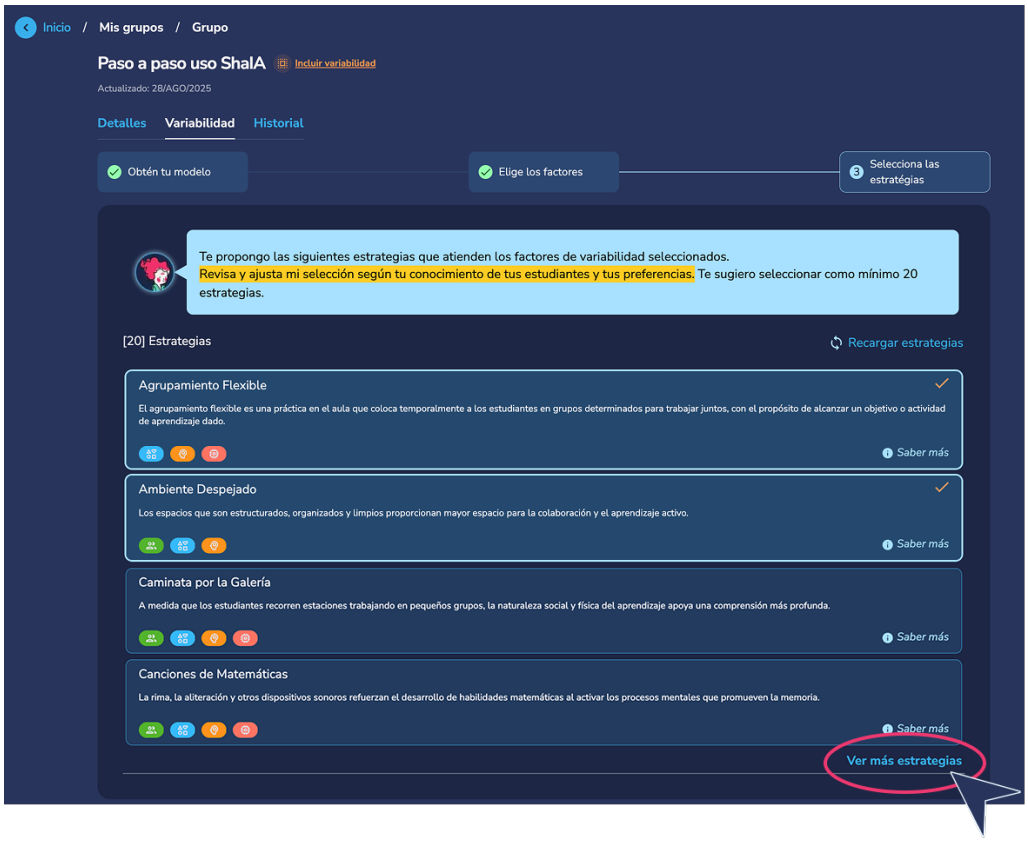
Appendix Figure A.4 presents *ShaIA*'s interface in which teachers select the factors they consider most relevant for their course. *ShaIA* provides an automatic selection of factors based on the information the teacher entered, but teachers can modify these selections based on their own professional 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 *Mentu* and *YouCubed*, these resources were translated into Spanish for the first time and made available to teachers participating in the intervention.

[3] A 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 *Shaia*
Pattern conversation	• 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	**Prioritised learning:** Learning focused on generalisation and algebraic thinking, typically linked to the metric–spatial component. **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. **Components:** Focuses on the metric–spatial component by working with visual representations, figure growth, and geometric patterns.	Example 1
Numerical conversation	• 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	**Prioritised learning:** Contributes to learning linked to the numerical–variational component, strengthening numerical sense and relationships between operations. **Mathematical thinking skills:** Develops communication, reasoning, and procedural practice through numerical flexibility. **Components:** Focuses primarily on the numerical–variational component.	Example 1
Mathematical experience	• 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	**Prioritised learning:** Allows explicit selection of prioritised learning and evidence as the central objective. **Mathematical thinking skills:** Integrates communication, reasoning, problem solving, and modelling. **Components:** Can address numerical–variational, metric–spatial and random components in an integrated or isolated manner.	Example 1
Worksheet	• Invite students to create graphic representations • Encourage collaborative work • Connect mathematical procedures and big ideas • Strategies for including variability	**Prioritised learning:** Allows you to set objectives and evidence directly associated with the SED's prioritised learning. **Mathematical thinking skills:** Promotes reasoning, communication, modelling, and problem solving. **Components:** The worksheet adapts to the prioritised learning selected by the teacher.	Example 1 Example 2 Example 3
Feedback practice	• Feedback and growth mindset messages • Error management • Recognition of variability factor • Strategies for including variability • Making students' thinking visible	**Prioritised learning:** Enables teachers to offer formative feedback connected to prioritised learning. **Mathematical thinking skills:** Reinforces mathematical communication and reasoning. **Components:** Adapts to the learning component worked on in class.	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:

$$\text{Visit 1} \rightarrow \text{Visit 2} \rightarrow \text{Workshop 1} \rightarrow \text{Visit 3} \rightarrow \text{Workshop 2} \rightarrow \text{Workshop 3} \rightarrow \text{Visit 4}.$$

   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.

Appendix Table A.1: Balance Across Schools In and Out of the Midline Sample (School-Level Outcomes)

	Observations		Average		Difference		Treated only		Controls only	
	In	Out	In	Out	Coef.	p-val.	Coef.	p-val.	Coef.	p-val.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Number of Teachers	29	129	89.10	84.74	4.36	0.58	1.22	0.92	6.91	0.55
Number of Teachers Sec. School	29	129	44.00	41.69	2.31	0.60	-1.15	0.85	5.11	0.44
Number of Math Teachers	29	129	6.21	5.40	0.81	0.25	0.63	0.54	0.81	0.42
Number of Math Teachers Sec. School	29	129	6.10	5.32	0.79	0.26	0.81	0.44	0.63	0.52
Permanent (%)	29	129	0.88	0.86	0.02	0.16	0.02	0.35	0.02	0.08*
Female (%)	29	129	0.67	0.67	-0.00	0.81	-0.02	0.43	0.01	0.32
Hired Under New Regime–2277 (%)	29	129	0.22	0.24	-0.02	0.45	0.01	0.73	-0.03	0.29
Secondary School (%)	29	129	0.49	0.49	-0.00	0.78	-0.02	0.22	0.01	0.41
Math (%)	29	129	0.07	0.06	0.00	0.29	0.01	0.26	0.00	0.77
Math Teachers in Sec. School (%)	29	129	0.14	0.13	0.01	0.19	0.03	0.01**	-0.00	0.70
Years Working in School	29	129	9.23	9.68	-0.45	0.52	-0.57	0.63	0.02	0.98
Age	29	129	48.36	48.72	-0.36	0.60	-0.17	0.89	-0.21	0.74
Number of Students	29	128	643.83	589.59	54.23	0.45	39.14	0.68	71.56	0.52
Number of Students in 6th Grade	29	128	173.28	161.48	11.79	0.53	13.86	0.61	10.94	0.70
Number of Students in 7th Grade	29	128	168.31	151.67	16.64	0.37	18.52	0.47	15.85	0.58
Number of Students in 8th Grade	29	128	154.62	143.73	10.89	0.55	2.53	0.91	19.88	0.49
Number of Students in 9th Grade	29	128	147.62	132.70	14.92	0.38	4.23	0.84	24.89	0.36
Students in Rural Schools (%)	29	128	0.03	0.01	0.03	0.45	-0.02	0.33	0.07	0.31
Students in Morning Session (%)	29	128	0.48	0.38	0.11	0.09*	0.18	0.08*	0.04	0.61
Students in Afternoon Session (%)	29	128	0.30	0.29	0.00	0.92	-0.03	0.70	0.04	0.50
Students in All-day Session (%)	29	128	0.22	0.33	-0.11	0.20	-0.15	0.26	-0.08	0.52
Low-Income Students (%)	29	128	0.26	0.29	-0.02	0.50	-0.10	0.02**	0.03	0.60
Age	29	128	14.47	14.48	-0.00	0.93	-0.07	0.24	0.04	0.54
Migrant Students (%)	29	128	0.13	0.12	0.01	0.57	-0.00	0.87	0.02	0.35
Students with Disabilities (%)	29	128	0.03	0.04	-0.01	0.20	-0.01	0.45	-0.01	0.40
Female Students (%)	29	128	0.49	0.49	-0.01	0.55	0.00	0.71	-0.01	0.51
Number of Sections	29	128	11.97	12.48	-0.51	0.54	-0.81	0.47	-0.31	0.81
Number of Sections in 6th Grade	29	128	3.14	3.31	-0.17	0.43	-0.08	0.81	-0.26	0.46
Number of Sections in 7th Grade	29	128	3.10	3.18	-0.08	0.71	-0.00	1.00	-0.17	0.60
Number of Sections in 8th Grade	29	128	2.86	3.08	-0.22	0.32	-0.33	0.30	-0.17	0.63
Number of Sections in 9th Grade	29	128	2.90	2.91	-0.02	0.95	-0.36	0.22	0.28	0.54
Average Saber 11 2019	25	121	0.00	0.12	-0.12	0.01**	-0.01	0.94	-0.21	0.00***
Average Saber 11 2020	25	120	-0.03	0.06	-0.08	0.13	-0.01	0.94	-0.15	0.06*
Math Saber 11 2019	25	121	-0.00	0.11	-0.11	0.04**	-0.01	0.88	-0.18	0.02**
Math Saber 11 2020	25	120	0.01	0.07	-0.06	0.27	0.02	0.78	-0.12	0.10
Missing Saber 11 (%)	29	129	0.14	0.06	0.08	0.27	0.15	0.18	0.00	0.96

Note: The sample includes the 69 schools observed at baseline, of which 29 were observed again at midline. *In* refers to schools observed at midline. *Out* refers to schools observed only at baseline. Columns (5)-(6) report the difference between the two groups (regression of each variable on a midline indicator). Columns (7)-(8) and (9)-(10) repeat this comparison restricting the sample to treated and control schools, respectively. Standard errors are heteroskedasticity-robust. *** p<0.01, ** p<0.05, * p<0.1.

Appendix Table A.2: 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. (1)	Con. (2)	Treat. (3)	Con. (4)	Unadj. (5)	Adj. (6)	Treat. (7)	Con. (8)	Treat. (9)	Con. (10)	Unadj. (11)	Adj. (12)	Treat. (13)	Con. (14)	Treat. (15)	Con. (16)	Unadj. (17)	Adj. (18)
A) Teachers																		
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*
B) Students																		
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
C) Classrooms																		
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
C) Test Scores																		
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

Appendix Table A.3: Balance Across Experimental Groups at Baseline

	Universe						Dropping Switchers						Dropping Declined					
	Observations		Average		P-Value		Observations		Average		P-Value		Observations		Average		P-Value	
	Treat. (1)	Con. (2)	Treat. (3)	Con. (4)	Unadj. (5)	Adj. (6)	Treat. (7)	Con. (8)	Treat. (9)	Con. (10)	Unadj. (11)	Adj. (12)	Treat. (13)	Con. (14)	Treat. (15)	Con. (16)	Unadj. (17)	Adj. (18)
A) Students																		
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*
B) Teacher Characteristics																		
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
C) Use of Digital Technologies																		
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
D) Teacher Practices																		
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
E) Perceptions in the Use of AI																		
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 ShaIA?	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

Appendix Table A.4: Baseline School Outcomes in Midline Sample of Schools

	Dropping Switchers						Pure Sample					
	Observations		Average		P-Value		Observations		Average		P-Value	
	Treat. (1)	Con. (2)	Treat. (3)	Con. (4)	Unadj. (5)	Adj. (6)	Treat. (7)	Con. (8)	Treat. (9)	Con. (10)	Unadj. (11)	Adj. (12)
A) Teachers												
Number of Teachers	14	14	90.071	91.071	0.947	0.888	11	9	89.636	94.778	0.800	.
Number of Teachers Sec. School	14	14	43.429	46.143	0.746	0.945	11	9	43.182	49.111	0.611	.
Number of Math Teachers	14	14	6.571	6.071	0.718	0.798	11	9	6.818	6.667	0.937	.
Number of Math Teachers Sec. School	14	14	6.571	5.857	0.601	0.798	11	9	6.818	6.444	0.843	.
Permanent (%)	14	14	0.880	0.896	0.453	0.979	11	9	0.890	0.906	0.488	.
Female (%)	14	14	0.664	0.686	0.341	0.624	11	9	0.676	0.678	0.937	.
Hired Under New Regime–2277 (%)	14	14	0.231	0.220	0.811	0.434	11	9	0.247	0.215	0.543	.
Secondary School (%)	14	14	0.479	0.500	0.395	0.642	11	9	0.475	0.507	0.214	.
Math (%)	14	14	0.071	0.065	0.425	0.900	11	9	0.072	0.068	0.637	.
Math Teachers in Sec. School (%)	14	14	0.150	0.123	0.075	0.754	11	9	0.153	0.126	0.179	.
Years Working in School	14	14	9.090	9.951	0.511	0.415	11	9	9.593	9.876	0.857	.
Age	14	14	48.624	48.753	0.908	0.461	11	9	49.073	49.014	0.961	.
B) Students												
Number of Students	14	14	649.500	662.000	0.927	0.777	11	9	652.000	734.889	0.661	.
Number of Students in 6th Grade	14	14	180.071	173.429	0.850	0.775	11	9	182.091	186.667	0.925	.
Number of Students in 7th Grade	14	14	174.643	168.429	0.862	0.837	11	9	178.364	185.444	0.887	.
Number of Students in 8th Grade	14	14	151.214	163.643	0.719	0.587	11	9	149.455	187.444	0.433	.
Number of Students in 9th Grade	14	14	143.571	156.500	0.691	0.853	11	9	142.091	175.333	0.451	.
Students in Rural Schools (%)	14	14	0.000	0.072	0.327	0.630	11	9	0.000	0.000	0.162	.
Students in Morning Session (%)	14	14	0.576	0.422	0.157	0.326	11	9	0.530	0.356	0.215	.
Students in Afternoon Session (%)	14	14	0.281	0.331	0.536	0.579	11	9	0.288	0.280	0.943	.
Students in All-day Session (%)	14	14	0.143	0.247	0.485	0.630	11	9	0.182	0.363	0.381	.
Low-Income Students (%)	14	14	0.230	0.297	0.289	0.860	11	9	0.239	0.256	0.823	.
Age	14	14	14.468	14.489	0.776	0.675	11	9	14.467	14.491	0.810	.
Migrant Students (%)	14	14	0.123	0.136	0.537	0.878	11	9	0.128	0.116	0.641	.
Students with Disabilities (%)	14	14	0.030	0.033	0.788	0.888	11	9	0.027	0.026	0.847	.
Female Students (%)	14	14	0.482	0.492	0.444	0.928	11	9	0.485	0.495	0.563	.
C) Classrooms												
Number of Sections	14	14	12.143	12.071	0.961	0.461	11	9	11.909	13.556	0.377	.
Number of Sections in 6th Grade	14	14	3.286	3.071	0.574	0.423	11	9	3.273	3.333	0.905	.
Number of Sections in 7th Grade	14	14	3.286	3.000	0.383	0.423	11	9	3.273	3.333	0.880	.
Number of Sections in 8th Grade	14	14	2.929	2.857	0.853	0.464	11	9	2.818	3.333	0.283	.
Number of Sections in 9th Grade	14	14	2.714	3.143	0.394	1.000	11	9	2.636	3.556	0.173	.
C) Test Scores												
Average Saber 11 2019	12	13	0.082	-0.073	0.072	0.810	10	8	0.074	-0.081	0.152	.
Average Saber 11 2020	12	13	0.031	-0.085	0.281	0.349	10	8	0.024	-0.087	0.359	.
Math Saber 11 2019	12	13	0.068	-0.064	0.156	0.048*	10	8	0.061	-0.079	0.239	.
Math Saber 11 2020	12	13	0.067	-0.051	0.239	0.874	10	8	0.062	-0.091	0.185	.
Missing Saber 11 (%)	14	14	0.143	0.071	0.558	0.423	11	9	0.091	0.111	0.889	.

Appendix Table A.5: Baseline Outcomes Across Experimental Groups in Midline Sample

	Dropping Switchers						Pure Sample					
	Observations		Average		P-Value		Observations		Average		P-Value	
	Treat. (1)	Con. (2)	Treat. (3)	Con. (4)	Unadj. (5)	Adj. (6)	Treat. (7)	Con. (8)	Treat. (9)	Con. (10)	Unadj. (11)	Adj. (12)
A) Students												
Baseline Test Score	1,154	1,210	32.04	32.32	0.85	0.04*	925	758	31.88	33.37	0.44	0.00*
Baseline Test Score (σ)	1,154	1,210	0.01	0.03	0.85	0.04*	925	758	-0.00	0.10	0.44	0.00*
Females	1,149	1,205	0.49	0.49	0.98	0.93	922	754	0.49	0.48	0.94	0.00*
Migrants	1,154	1,210	0.34	0.32	0.84	0.09	925	758	0.33	0.29	0.60	0.00*
International Migrants	1,154	1,210	0.10	0.10	0.96	0.06	925	758	0.09	0.08	0.72	0.00*
Wealth PC Index	1,154	1,210	0.12	-0.07	0.23	0.09	925	758	0.14	0.01	0.49	0.00*
Use of AI	1,153	1,210	0.82	0.87	0.12	0.24	924	758	0.83	0.86	0.43	0.00*
Advance Knowledge of AI	1,154	1,210	0.28	0.33	0.06	0.03*	925	758	0.29	0.34	0.15	0.00*
Math Learning Perceptions	1,140	1,196	-0.03	-0.02	0.90	0.09	911	751	0.06	-0.05	0.23	0.00*
Math Class Perceptions	1,141	1,199	-0.02	0.02	0.60	0.14	915	750	0.02	0.02	0.99	0.00*
Teacher Perceptions	1,122	1,175	-0.07	0.08	0.10	0.03*	901	736	0.04	0.08	0.32	0.00*
Expected to Complete College	1,154	1,210	0.80	0.82	0.45	0.01*	925	758	0.81	0.83	0.51	0.00*
Expected to Continue in STEM	1,154	1,210	0.25	0.22	0.09	1.00	925	758	0.26	0.20	0.05	0.00*
AI can be helpful in the future	1,154	1,210	0.34	0.32	0.45	0.25	925	758	0.35	0.30	0.06	0.00*
B) Teacher Characteristics												
Female	46	74	0.43	0.36	0.40	0.00*	34	48	0.41	0.42	0.96	1.00
Age	46	74	45.39	46.81	0.42	0.00*	34	48	47.03	47.04	1.00	0.52
High School or Less	46	74	0.09	0.08	0.91	0.02*	34	48	0.12	0.04	0.20	1.00
University Degree	46	74	0.07	0.05	0.82	0.06	34	48	0.00	0.06	0.02*	1.00
Postgraduate	46	74	0.85	0.86	0.75	0.57	34	48	0.88	0.90	0.82	1.00
Teaching Experience	46	74	19.46	20.55	0.53	0.00*	34	48	21.53	20.08	0.37	0.00*
Work Experience	46	74	22.41	24.61	0.21	0.02*	34	48	24.53	24.71	0.90	0.00*
C) Use of Digital Technologies												
Digital Resources at home (Index)	45	73	0.27	-0.03	0.29	0.46	34	48	0.33	-0.08	0.28	0.00*
Digital Resources at work (Index)	43	66	0.28	-0.13	0.23	0.97	31	43	0.35	-0.10	0.34	0.00*
First computer use age > 20	46	74	0.26	0.28	0.74	0.00*	34	48	0.32	0.27	0.47	0.00*
First internet use age > 20	46	74	0.39	0.42	0.74	0.22	34	48	0.44	0.44	0.97	1.00
Internet > 2h per day	46	74	0.72	0.70	0.88	0.05	34	48	0.76	0.65	0.36	1.00
Internet at School > 2h per day	46	74	0.37	0.36	0.96	0.09	34	48	0.41	0.29	0.29	0.00*
Uses AI at Work	46	74	0.72	0.73	0.89	0.26	34	48	0.71	0.73	0.85	1.00
Knowledge of AI	46	74	0.87	0.84	0.60	0.00*	34	48	0.82	0.83	0.90	1.00
Knowledge of AI to teach	46	74	0.70	0.81	0.20	0.68	34	48	0.65	0.81	0.16	1.00
Uses AI to Teach	46	74	0.46	0.46	0.98	0.00*	34	48	0.38	0.46	0.30	0.63
Digital Use (Index)	41	66	0.55	-0.08	0.07	0.20	30	43	0.53	0.05	0.20	0.00*
Digital Use In-Class (Index)	43	72	0.01	-0.10	0.67	1.00	31	46	0.17	0.05	0.66	0.00*
Perc. Digital Use in Class (Index)	45	70	0.33	0.34	0.96	0.06	33	46	0.10	0.27	0.58	0.00*
D) Teacher Practices												
Last class prepared > 1h	46	74	0.35	0.38	0.71	0.69	34	48	0.26	0.33	0.44	0.00*
Teacher Aptitude (Index)	46	68	-0.06	0.07	0.74	0.17	34	42	-0.17	0.07	0.64	0.00*
Class time: Lecture (%)	46	74	26.59	27.77	0.52	0.95	34	48	26.26	28.02	0.43	0.00*
Class time: Practical Exercises (%)	46	74	22.39	22.32	0.97	0.31	34	48	23.09	22.44	0.77	0.00*
Class time: Group Work (%)	46	74	18.04	18.22	0.93	0.23	34	48	18.82	18.19	0.82	0.00*
Class time: Discussion (%)	46	74	11.83	10.50	0.20	0.00*	34	48	11.15	10.56	0.66	0.00*
Class time: Use of Technology (%)	46	74	6.33	5.03	0.30	0.63	34	48	6.50	4.31	0.15	0.00*
Class time: Assessments (%)	46	74	10.13	11.20	0.41	0.55	34	48	9.94	11.40	0.23	0.00*
Class time: Other (%)	46	74	4.70	4.96	0.77	0.27	34	48	4.24	5.08	0.44	0.00*
Teacher Own Perceptions (Index)	44	71	-0.15	0.02	0.71	0.77	32	46	-0.10	0.01	0.86	0.00*
E) Perceptions in the Use of AI												
Perceptions on AI to teach (Index)	40	66	-0.22	0.05	0.60	0.48	28	41	-0.36	-0.07	0.67	0.00*
AI is useful to teach	46	74	0.57	0.38	0.09	0.74	34	48	0.62	0.40	0.06	0.00*
AI helps: Class prep.	46	74	1.00	0.92	0.01*	0.01*	34	48	1.00	0.92	0.05	1.00
AI helps: Resources	46	74	0.98	0.91	0.12	0.01*	34	48	0.97	0.90	0.25	1.00
AI helps: Track Learning	46	74	0.76	0.61	0.16	0.58	34	48	0.85	0.58	0.02*	1.00
AI helps: Personalized Teaching	46	74	0.93	0.78	0.05	0.41	34	48	0.94	0.77	0.11	1.00
Last Year: Seminar on AI (%)	46	74	0.22	0.23	0.86	0.25	34	48	0.24	0.25	0.88	0.00*
Last Year: Course on AI (%)	46	74	0.24	0.26	0.78	0.92	34	48	0.24	0.25	0.85	1.00
Last Year: Use an AI Tool (%)	46	74	0.65	0.64	0.87	0.77	34	48	0.68	0.67	0.94	0.00*
Last Year: Use AI Tool to Teach (%)	46	74	0.48	0.50	0.82	0.89	34	48	0.47	0.48	0.94	1.00
Last Year: Research on AI (%)	46	74	0.48	0.46	0.86	0.53	34	48	0.47	0.48	0.95	0.00*
Last Year: Discussed AI (%)	46	74	0.61	0.50	0.22	0.60	34	48	0.62	0.48	0.23	0.00*
Last Year: Other AI Training (%)	46	74	0.26	0.22	0.61	0.31	34	48	0.29	0.23	0.58	0.00*
¿Conoce la herramienta ShaIA?	45	72	0.07	0.01	0.18	0.23	33	47	0.09	0.02	0.19	.

Appendix Table A.6: Baseline Test Score Differences by Gender and Migrant Status

		Migrant Status					Gender		
	Obs	Locals	Migrants		Differences		Female	Male	Difference
			Internal	External	(2) vs (3)	(2) vs (4)			(7) vs (8)
	(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