Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 26-1496

From Pilot to Policy: Experimental Evidence from Scaling Online Tutoring

María Calle
Seguridad Social y Migraciones

Lucas Gortazar
Esade Business School and World Bank

Maria Hernandez-de-Benito
CUNEF Universidad

Claudia Hupkau
CUNEF Universidad

Teresa Molina-Millán
University of Alicante

Antonio Roldán-Monés
IE University

We study a randomized controlled trial of an online mathematics tutoring program scaled from a successful pilot and implemented entirely by regional education authorities in Spain, using interim public-school teachers rather than specially recruited tutors. Assignment to tutoring increased end-of- year grades by 0.15σ and standardized math test scores by 0.11σ — approximately one-third of pilot effects. An experimental socio-emotional module improved affective outcomes but not academic gains. Results provide realistic benchmarks for the "voltage drop" policymakers should anticipate when scaling evidence-based interventions.

VERSION: June 2026

Suggested citation: Calle, María, Lucas Gortazar, Maria Hernandez-de-Benito, Claudia Hupkau, Teresa Molina-Millán, and Antonio Roldán-Monés. (2026). From Pilot to Policy: Experimental Evidence from Scaling Online Tutoring. (EdWorkingPaper: 26-1496). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/a7nb-1452

From Pilot to Policy: Experimental Evidence from Scaling Online Tutoring*

María Calle† Lucas Gortazar‡ Maria Hernandez-de-Benito§
Claudia Hupkau¶ Teresa Molina-Millán‖ Antonio Roldán-Monés**

May 25, 2026

Abstract

We study a randomized controlled trial of an online mathematics tutoring program scaled from a successful pilot and implemented entirely by regional education authorities in Spain, using interim public-school teachers rather than specially recruited tutors. Assignment to tutoring increased end-of-year grades by 0.15σ and standardized math test scores by 0.11σ — approximately one-third of pilot effects. An experimental socio-emotional module improved affective outcomes but not academic gains. Results provide realistic benchmarks for the "voltage drop" policymakers should anticipate when scaling evidence-based interventions.

Keywords: Online tutoring, scale-up, RCT, government implementation, mathematics achievement, socio-emotional learning

JEL Classification: C93, I21, I24, I28, J45.

*We are grateful to seminar and conference participants for valuable comments and suggestions. We thank the Government of Navarre's Department of Education for taking full responsibility for the implementation and administration of the scale-up, and to the Ministry of Inclusion, Social Security and Migration for facilitating access to the administrative data used in this study and funding. Claudia Hupkau acknowledges financial support from the Ministerio de Ciencia e Innovación through grants PID2022-140206OB-I00 and RYC2023-044087-I. Teresa Molina-Millán acknowledges financial support from Generalitat Valenciana through the plan GenT program (CDEIGENT/2020/016). All errors remain our own.

†Ministerio de Inclusión, Seguridad Social y Migraciones.

‡EsadeEcPol, Esade Business School and World Bank. Email: lucas.gortazar@esade.edu

§CUNEF Universidad. Email: m.hernandezdebenito@cunef.edu.

¶CUNEF Universidad and Centre for Economic Performance, London School of Economics. Email: claudia.hupkau@cunef.edu.

‖University of Alicante. Email: teresa.molina@ua.es.

**Department of Economics, IE University. Email: Antonio.roldan@ie.edu.

1 Introduction

Personalized, high-dosage tutoring has long been recognized as one of the most effective educational interventions, with recent meta-analyses documenting effect sizes often exceeding 0.3 standard deviations on academic outcomes (Nickow et al., 2024; Kraft and Falken, 2021). Technological advances have made high-quality tutoring increasingly accessible through online platforms, expanding reach while reducing costs (Carlana and La Ferrara, 2025; Kraft and Lovison, 2025; Gortazar et al., 2024; Kraft et al., 2022; Hardt et al., 2023; Escueta et al., 2020). Yet, despite accumulating evidence from carefully designed pilots, fundamental questions remain about whether these promising results can be replicated when programs scale beyond their original contexts and are implemented by actors other than their designers (Banerjee et al., 2017a; Justman, 2018).

The challenge of scaling effective educational interventions is well-documented. Programs demonstrating strong impacts in small-scale trials often show attenuated effects when expanded (Al-Ubaydli et al., 2017; Bold et al., 2018; Kraft et al., 2024a; Agostinelli et al., 2025)—a phenomenon called "voltage drop" (List, 2022). Multiple factors can contribute to this decline: dilution of program quality, reduced implementation fidelity, or the loss of founder-specific expertise and commitment. The risk may intensify when government agencies assume implementation responsibility, as public sector actors typically suffer from political economy constraints, or lack the operational flexibility and specialized knowledge of the program designers. Moreover, scaling often requires either hiring professional administrators or reallocating existing staff, risking increased cost or a potential reduction in the quality of program inputs (Al-Ubaydli et al., 2017). Understanding whether and how effective interventions survive the transition from pilot to scale, from founders to government implementers, and from external contractors to existing public sector staff has become a central concern in education policy across the globe.

This paper contributes to this debate by studying the scale-up of an online mathematics tutoring program for vulnerable students (Menttores) implemented in 2021 in Spain. Gortazar et al. (2024) documented significant positive impacts of Menttores, with gains of 0.26 standard deviations on standardized tests and 0.49 standard deviations on mathematics grades. The pilot program, delivered in groups of two students by qualified mathematics teachers over eight weeks, also reduced grade retention and increased students' educational aspirations and self-reported effort. Although encouraging, the pilot study covered only 375 students in a limited number of innovation-prone urban schools, was operated under direct founder oversight with intensive quality control, and was delivered by implementation experts and teachers recruited through a highly competitive selection process. This leaves critical questions unanswered: Can these effects be reproduced when the program scales, confronting real implementation and political economy constraints that governments typically face? Can government education authorities, operating without the founders' involvement, successfully run such interventions? And crucially, are there ways for governments to leverage existing underutilized staff rather than requiring new hires?

To address these questions, we designed a large-scale randomized evaluation in which the program was implemented and administered entirely by regional education authorities in Navarre, Spain. As part of the evaluation, 1,344 students were randomized across 46 schools—representing 15% of all schools in the region—to receive online mathematics tutoring (971 students) or a control condition (373 students), tracking academic outcomes through administrative records and standardized assessments, as well as socio-emotional outcomes through student surveys. The scale-up was implemented through multiple treatment arms that experimentally varied (i) tutoring group size (two vs. three students per tutor), and (ii) whether tutors received enhanced socio-emotional training alongside the standard tutoring model.

This implementation represents a demanding test of scalability, transferability, and sustainability on three key dimensions. First, the program expanded to approximately 3.6 times its pilot size, including both urban and rural areas. Second, operational control shifted completely from founders to government officials who had no involvement in the original design or previous experience in RCT implementation. Third, and perhaps most critically for policy, tutoring was delivered exclusively by interim teachers (interinos)—qualified educators already on the government registry who face employment gaps between permanent assignments and represent a readily available, underutilized resource for education authorities. This staffing transition is particularly demanding: the original pilot employed tutors recruited through Empieza por Educar, a Spanish NGO (modeled on Teach for America) that rigorously screens candidates on motivation, empathy, and commitment to working in disadvantaged schools alongside academic credentials. Moving from this elite pool of highly dedicated educators to a readily available government workforce represents a demanding test of whether tutoring effects persist under realistic implementation conditions.

Our findings provide qualified evidence on the prospects for scaling effective tutoring programs under real-world government implementation. We document modest but significant positive effects on mathematics outcomes: assignment to tutoring increased math grades by 0.15σ ($p < 0.01$) and standardized math test scores by 0.11σ ($p < 0.10$), with no spillover effects on overall GPA. These estimates represent approximately 30-42% of the pilot effect sizes, confirming the existence of significant voltage drop but demonstrating that meaningful impacts survive scale-up and transfer to government control. For socio-emotional outcomes—though with the caveat that endline survey attrition and non-response were substantial—we find significant improvements in math self-efficacy and reduced anxiety (0.22σ, $p < 0.05$), with an experimental socio-emotional learning (SEL) module producing additional benefits for affective outcomes–including interest in mathematics, life satisfaction, and interest in other subjects–without enhancing academic gains.

A central challenge in scaling tutoring is balancing effectiveness with cost-efficiency. While smaller student-tutor ratios are generally presumed more effective, they also limit program reach under fixed budgets–a binding constraint for government-led implementation. Through experimental variation in group size (2:1 versus 3:1 student-tutor ratios), we test whether increases in group size compromise outcomes. We find statistically indistinguishable effects across these ratios, suggesting that programs could potentially reduce per-student costs by approximately a third by serving three rather than two students per tutor without

sacrificing effectiveness. This finding addresses one of Kraft and Falken (2021)'s central scaling challenges—cost constraints—by identifying an efficiency gain that improves fiscal sustainability without compromising program quality.

We also find that treatment effects are relatively uniform across student subgroups, but the intervention was more effective in enhancing math outcomes among younger (primary school) students and those in urban areas, as well as higher ability students as measured by their baseline performance, mirroring findings from the pilot (Gortazar et al., 2024) and other tutoring programs (Guryan et al., 2023).

Beyond demonstrating that tutoring effects can persist at scale under government implementation, this study offers important insights for education policy. First, the reliance on interim teachers addresses a key constraint facing many education systems: how to deliver intensive interventions without large budget increases or new hiring. These educators, already credentialed and on the government's teacher registry, represent latent capacity within existing government structures. Our results suggest that mobilizing this underutilized workforce can yield cost-effective improvements in student outcomes, making large-scale tutoring programs more fiscally sustainable than models requiring external recruitment. Second, the use of interim teachers helps mitigate political-economy constraints associated with scaling tutoring programs. In particular, reliance on volunteer labor or ad hoc private hiring can generate resistance within government systems, whereas interim teachers are already integrated into existing institutional structures. Third, delivering the program entirely online not only reduces tutoring costs—such as transportation and related logistical expenses—but also enhances scalability by relaxing teacher supply constraints, particularly in sparsely populated or remote rural areas.

The observed voltage drop, while substantial, may represent an acceptable trade-off given the reduced marginal cost and enhanced scalability of leveraging existing staff. Moreover, implementation challenges—including incomplete session attendance and the absence of founder oversight—provide realistic benchmarks for what policymakers can expect when adopting tutoring programs at scale.

This study contributes to several literatures. First, we provide rigorous evidence on a central question of external validity: whether tutoring effects survive the transition from pilot to scale under real-world government implementation. A large body of work documents that tutoring is among the most effective educational interventions, with meta-analyses finding average effects of 0.2–0.4σ (Nickow et al., 2024; Kraft and Falken, 2021). Yet, the few studies that have addressed tutoring programs at scale reveal a more sobering picture. Kraft et al. (2024b) document how Nashville's "Accelerating Scholars" program—scaled from 132 to over 6,800 students—produced modest effects of 0.04–0.09σ on reading, while facing substantial implementation challenges that forced a pivot from online to predominantly in-person delivery. More broadly, the "voltage drop" phenomenon is well-documented across educational interventions: programs showing strong impacts in carefully controlled trials often exhibit attenuated effects when expanded (List, 2022; Al-Ubaydli et al., 2017; Vivalt, 2020; Bold et al., 2018; Banerjee et al., 2017b; Muralidharan and Niehaus, 2017). However, many of these studies compare scaled programs that differed substantially from the pilot

interventions. Our study directly addresses this gap by experimentally evaluating a program that was explicitly designed as a scale-up of a successful pilot, allowing us to quantify precisely voltage drop rather than inferring it from cross-program comparisons. We also extend the predominantly US-focused evidence to the European context, where education systems, teacher labor markets, and government administrative capacity differ substantially.

Second, we demonstrate a novel staffing and delivery model that addresses the two most binding constraints on tutoring scale-up: tutor recruitment and physical infrastructure. Ambitious proposals for national tutoring programs face fundamental supply-side challenges (Kraft and Falken, 2021). Volunteer-based models, while cost-effective, suffer from high turnover and variable engagement (Kraft et al., 2022; Carlana and La Ferrara, 2025). Professional tutoring requires either costly new hires or reallocating existing staff, potentially reducing instructional quality elsewhere (Guryan et al., 2023). Nashville's experience illustrates these tensions: the program shifted from volunteers to teachers, but the move to in-person delivery during planning periods created scheduling bottlenecks and sustainability concerns. Our approach—pairing online delivery with interim teachers—offers a potential solution. Online tutoring eliminates physical space constraints and enables flexible scheduling, while drawing from existing credentialed staff avoids recruitment bottlenecks and ensures professional standards. In political economy terms, utilizing remunerated government staff for program delivery might be more acceptable to key education stakeholders, including teacher unions, as compensating educators for additional instructional work helps maintain professional standards and avoids potential conflicts that can arise when volunteers are positioned as substitutes for remunerated professionals in core educational functions. This model may be particularly relevant in contexts with significant pools of qualified but underemployed teachers, a common feature of education systems with competitive civil service examinations. The scale of this underutilized buffer varies across countries, but is far from negligible even in high-income settings. OECD data show that roughly 17% of primary and secondary teachers in OECD countries hold fixed-term contracts, with the share reaching 27% in Spain, 26% in the US, and over 20% in Chile and Italy (OECD, 2019). Part-time arrangements are similarly widespread: in Mexico and the Netherlands, over half of lower-secondary teachers report working part-time, and the share increased significantly in ten OECD countries between 2013 and 2018 (OECD, 2019). This buffer is likely larger still in middle-income countries with competitive civil service systems.

Third, we contribute to the growing literature on online teaching effectiveness. Carlana and La Ferrara (2025) show that intensive volunteer-led online tutoring during and after the pandemic improved math performance by approximately 0.2σ in Italy. In the US context, evidence is more mixed: Kraft et al. (2022) find positive but imprecise effects from volunteer online tutoring, while Bettinger et al. (2017) document negative effects of fully online coursework, suggesting that technology complements rather than substitutes for human instruction; similarly, Ajzenman et al. (2024) find that replacing trained human promoters with rule-based chatbots attenuates impacts in a scalable behavioral intervention aimed at motivating high school students to enroll in education majors. Our findings indicate that online tutoring can be sustained at scale

by governments when delivered by compensated professionals with adequate technical infrastructure.

Fourth, we provide experimental evidence on the relationship between academic and socio-emotional components in tutoring programs. Meta-analyses suggest that SEL programs can improve both non-cognitive skills and academic achievement (Durlak et al., 2011), and (Kraft and Falken, 2021), for instance, argue that integrating mentoring elements into tutoring may broaden student benefits. However, the optimal intensity of such integration is unclear. Brown et al. (2023) find that adding targeted SEL activities to remedial tutoring in Niger improves school grades but not standardized test scores, and paradoxically does not improve measured SEL outcomes. Our experimental design—randomly assigning tutors to receive enhanced SEL training through an additional 12-hour module—allows us to test whether intensifying the socio-emotional component yields additional benefits. We find that enhanced SEL training improves affective outcomes (math interest, life satisfaction) but does not generate additional academic gains, suggesting diminishing returns to SEL intensity beyond a foundational level. This result has practical implications for program design: marginal investments in additional SEL training may operate primarily through affective rather than academic pathways (Jackson, 2018; Algan and Huillery, 2025).

The remainder of the paper proceeds as follows. Section 2 describes the intervention design, experimental protocol, and recruitment. Section 3 describes data sources, and Section 4 describes the empirical strategy. Section 5 presents main results on academic and socio-emotional outcomes, including robustness checks for differential attrition. Section 6 discusses implications for educational policy and concludes.

2 Program design and implementation

In this section we discuss the background of the intervention, describe the program design, school, student and tutor recruitment, randomization and timeline.

2.1 Background

The intervention evaluated in this study builds on a successful pilot that paired external tutors with lower-secondary students in the regions of Madrid and Barcelona, Spain (Gortazar et al., 2024). The pilot program consisted of an eight-week, two-on-one online mathematics tutoring model, with three sessions per week delivered by external professional mathematics teachers hired specifically for the program. The initial intervention had a significant positive impact on standardized mathematics scores ($+0.26\sigma$) and end-of-year grades ($+0.49\sigma$), and it also increased students' aspirations—measured as the likelihood of stating they would like to pursue the academic track in upper secondary school—as well as self-reported effort at school.

In Spain, educational policy is devolved to its autonomous regions, which hold extensive powers over various aspects of the education system. These include the regulation of non-basic elements of the curriculum, the management of educational institutions, the development of region-specific content, and the allocation of public funding for education (European Committee of the Regions, 2023). As a result, regional authorities have the autonomy to design and implement additional support programs—like the one evaluated in this

study—for their students .

After the publication of the results from the pilot study, the government of Navarre expressed interest in scaling up the program through its Department for Education. Navarre is a region with more than 600,000 inhabitants and over 72,000 students enrolled in 311 primary and secondary schools (Gobierno de Navarra, Departamento de Educación, 2024).[1] The government of Navarre was fully responsible for the execution and implementation of the program through its Department for Education. Funding for the scale-up was provided by Spain's Ministry of Inclusion, Social Security and Migration under the umbrella of the Spanish Inclusion Policy Lab.[2]

Primary Education (Educación Primaria) spans from ages 6 to 12 (1st to 6th grade), and Compulsory Secondary Education (Educación Secundaria Obligatoria, or ESO), covers ages 12 to 16 (1st to 4th year of ESO). Upon completing ESO, students receive the Graduate of Secondary Education Certificate (Graduado en Educación Secundaria Obligatoria), qualifying them to pursue post-compulsory education.

After compulsory schooling, students can choose between two main pathways. The first is the academic track, leading to the Bachillerato, a two-year program typically undertaken between ages 16 and 18. The Bachillerato offers various specializations, such as Sciences and Technology, Humanities and Social Sciences, and Arts. Successful completion of the Bachillerato is a prerequisite for university admission, which also requires passing the university entrance examination known as the Evaluación para el Acceso a la Universidad (EvAU) or Selectividad. Alternatively, students may opt for vocational education through the Formación Profesional (FP) system. This pathway includes Intermediate Level Vocational Training (Grado Medio), generally lasting two years and combining theoretical instruction with practical experience.

2.2 Experimental Design

The program involved 1,344 students from 46 public schools across Navarre that met defined thresholds of poverty and vulnerability. Participating students ranged from 5th year of primary school through 2nd year of secondary school (grades 5-8). In the pre-intervention standardized mathematics test, the average score in both waves was 3 out of 10, reflecting the low baseline achievement levels of participating students. The intervention was delivered across two waves: spring (April-June) and fall (September-November) 2023, with 545 students (40%) participating in Wave 1 and 799 students (60%) in Wave 2.

The experimental design introduces variation across three dimensions to test specific hypotheses about scalability and program effectiveness. First, by comparing any tutoring condition to control, we estimate

[1] Of the 72,000 compulsory school-aged students attending primary and lower secondary schools, approximately 46,000 are enrolled in public schools and 26,000 in state-subsidized private schools (*escuelas concertadas*). Of the 311 primary and secondary schools, 227 are fully public, 84 are state-subsidized private, and 4 are fully private.

[2] The Inclusion Policy Lab is a national initiative launched by Spain's Ministry of Inclusion, Social Security and Migration (MISSM) to rigorously test and improve social inclusion policies. Funded through the European Union's Recovery and Resilience Facility, the Lab supported 32 pilot projects across Spain. Its central goal was to generate causal evidence on effective strategies for promoting the socio-labor inclusion of vulnerable populations. All projects were designed and evaluated using randomized controlled trials (RCTs), following a common framework developed in collaboration with academic partners, including CEMFI and J-PAL Europe. For additional details, see *Laboratorio de Políticas de Inclusión*, Ministerio de Inclusión, Seguridad Social y Migraciones (MISSM) https://www.inclusion.gob.es/web/policy-lab/laboratorio.

whether effects survive a 3.6-fold scale-up under full government implementation. Second, we randomly vary group size (two versus three students per tutor) to test whether modest increases in student-tutor ratios attenuate impacts while improving cost-effectiveness. Third, in Wave 2, we randomly assigned some tutors to receive enhanced socio-emotional training (detailed in Section 2.4), testing whether intensifying the non-academic component improves socio-emotional and academic outcomes following the limited SEL impacts observed in the pilot. Figure 1 summarizes the sample allocation across experimental conditions and waves.

The intervention closely followed the pilot design while adapting it along two main dimensions. First, to better accommodate school schedules and tutor availability, the scale-up maintained the same total instructional time as the pilot (180 minutes per week) but restructured delivery into two 90-minute sessions as opposed to three 60-minute sessions.[3]

Second, the key difference between pilot and scale-up was the tutor workforce (further described in Section 2.5). The pilot recruited tutors through Empieza por Educar, a selective program modeled on Teach for America that rigorously screens candidates on motivation, empathy, and commitment alongside academic credentials. The scale-up utilized interim teachers already registered in the government's *bolsa de interinos*—credentialed educators integrated into existing civil service structures who face employment gaps between assignments. This shift from highly selected, mission-driven educators to readily available government staff represents the core test of whether tutoring effects survive real-world government implementation.

2.3 Program implementation

All treatment groups participated in online mathematics tutorials given by teaching professionals, during extracurricular hours and over a period of eight weeks. Tutoring sessions lasted 90 minutes and were provided twice a week, for a total of 16 sessions. Each tutor had a maximum of six groups of students.

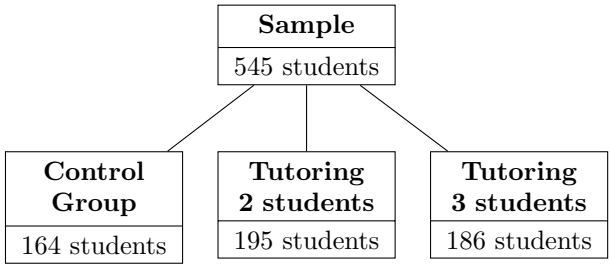
Sessions were delivered online from dedicated facilities at CREENA (Resource Centre for Educational Equity in Navarre) in Pamplona and Tudela, where tutors were provided technical support and adequate infrastructure. To support coordination with regular schooling, tutors traveled to participating schools each Friday to meet with students' classroom teachers. A dedicated tutor coordinator, drawn from the pool of contracted tutors, acted as liaison between students, classroom teachers, and tutors, providing organizational support, and covering sessions when necessary.

Tutors received a specific training of 15 hours in methodologies and digital tools, with a pedagogical component to teach classes effectively online. This is considered the "baseline" treatment and replicates the training provided in the pilot program (Gortazar et al., 2024). Tutors randomly assigned to the socio-emotional treatment arm received an additional 12 hours of training based on the Programa Laguntza for emotional well-being (see Section 2.5 for details).

[3]Existing evidence suggests total dosage matters more than session distribution for learning outcomes (Nickow et al., 2024).

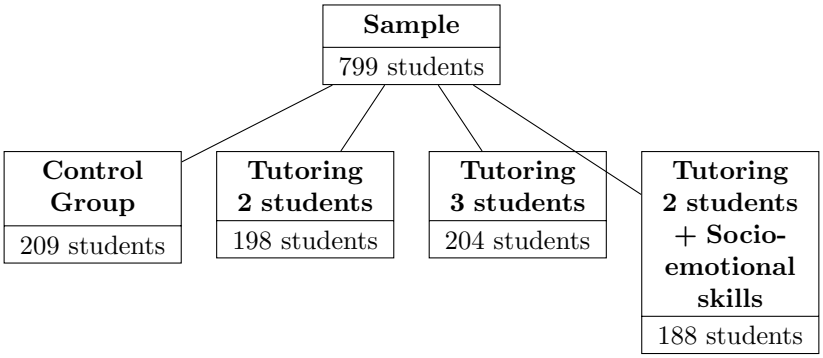
Figure 1: Sample allocation across experimental conditions and waves

(a) Wave 1



Stratification: school, grade, and sex.

(b) Wave 2



Stratification: school, grade, sex, and main language.

2.4 Recruitment of schools and participants

The Ministry of Education of Navarre targeted the intervention at schools meeting specific vulnerability criteria. To determine areas of high poverty, two indicators from the Navarre Observatory of Social Reality were considered: (1) the relative at-risk-of-poverty rate for children under 16 years of age, measured as the percentage of children living in households whose total annual equivalent income is below 60% of the median; and (2) the at-risk-of-severe-poverty rate for children under 16 years of age, measured as the percentage of children living in households whose total annual equivalent income is below 40% of the median. Schools identified as potential beneficiaries were those located in areas where the sum of both indicators was above a pre-defined threshold. This threshold was lowered in wave 2 to ensure a sufficiently large sample size.

After applying the eligibility criteria, 49 schools were identified in the first wave and 85 in the second wave as potential beneficiaries. The final selection included schools in both urban and rural areas. All eligible schools were invited to participate. Schools that agreed to participate (18 out of 49, or 37%, in wave 1 and 39 out of 85, or 46%, in wave 2) were subsequently responsible for informing students and families about the opportunity to enroll. After contacting families, thoroughly explaining the project, and obtaining informed consent, the sample of participating students was finalized. This process was followed in both waves.[4]

[4]In the first wave, as an incentive to participate, families were informed that students assigned to the control group would have the opportunity to receive tutoring during the fall term of the same year. In the recruitment process for the second wave, this incentive was not offered.

2.5 Recruitment, training and coordination of tutors

The tutoring program implemented by the regional government of Navarre relied on the existing pool of interim teachers listed in the *bolsa de interinos*—a registry of qualified individuals who had passed part or all of the competitive civil service exams (*oposiciones*) but had not secured a permanent post. Placement on the list follows a ranking system based on points awarded for exam performance and other merit-based criteria. Tutoring positions were offered to candidates on this list with no separate application process. The offers specified the key working conditions: full-time contracts for 8 to 10 weeks, with tutoring sessions scheduled in the afternoon. The salary and contractual conditions were equivalent to those of standard teaching posts. Importantly, candidates could decline the tutoring offer without incurring penalties that would affect their position in the registry.

Table A.1 presents the characteristics of the recruited tutors. Tutors were predominantly female (82% and 69% waves 1 and 2, respectively), of Spanish nationality (97%), and the mean age among tutors was 32–34 years. More than 90% of recruited tutors had previously worked as a teacher, and 85% had worked as math teachers. This is not surprising given that tutors were drawn from the existing pool of interim teachers. A substantial fraction had worked as tutors before (50% in wave 1 and 70% in wave 2).

Training was provided after appointment and during paid working hours. All tutors completed a 15-hour training course in basic digital tools for online teaching, designed and delivered by Navarre's Department for Education. In addition, tutors assigned to the subgroup focusing on socio-emotional competencies received 12 hours of supplementary training based on the *Programa Laguntza* (Gobierno de Navarra, Departamento de Educación, 2025) for emotional well-being. This training was delivered in four sessions, either in-person or online and is described in more detail below.

Training in socio-emotional competencies The *Programa Laguntza*, developed by the Department of Education of the Government of Navarra, is a school-based initiative aimed at enhancing emotional well-being and fostering a positive coexistence climate within the educational system. While the program was initially conceived as a means of addressing bullying and interpersonal conflicts in schools, it has since undergone a conceptual shift toward a broader framework of socio-emotional education (Noticias de Navarra, 2023).

Within the context of the online tutoring program, the training focused on three core areas: emotional awareness, interpersonal competence, and restorative practices. Teachers are first introduced to the conceptual foundations of emotional education, including the identification, understanding, and regulation of emotions in both students and themselves. This includes practical strategies for recognizing emotional cues, managing stress, and promoting emotional expression within pedagogical contexts. Second, the program emphasizes the development of empathy and active listening skills, enabling teachers to build more supportive and respectful relationships with students. Educators learn to facilitate classroom dialogue, manage group dynamics constructively, and reinforce prosocial behaviors. Third, training incorporates basic restorative practices aimed at conflict resolution and the reparation of relationships, including techniques for facilitating

mediated conversations, encouraging accountability, and restoring trust after interpersonal tension.

Overall, the Laguntza-informed training equips teachers with a set of pedagogical tools for integrating socio-emotional competence into their instructional practice, thereby creating a learning environment conducive to both academic and personal growth.

2.6 Randomization

After obtaining consent and finalizing the sample selection for the study, participants in the experiment were randomly assigned to either the control group or one of various treatment groups, as shown in Figure 1.[5] Randomization was done at the student level, employing a stratified approach to ensure balancing on certain characteristics relevant to outcome indicators and to control certain aspects of tutoring group student composition. Stratification took into account school, grade, and sex in the first wave, and main language (Spanish or Basque) was added in the second wave.

The strata based on school, grade, and main language were "strict," meaning each tutoring group had to consist of students who were from the same school and grade and who shared the same main language. Language was included as a stratification variable in the second wave because some schools operate with multiple linguistic models, where certain classes are taught in Basque and others in Spanish; thus, tutoring needed to be delivered in the students' language of instruction. In the first wave, this constraint was not incorporated into the stratification design because it was raised by the implementing counterpart only after randomization; language matching was therefore implemented ex-post. Gender serves as a stratification variable to ensure gender balance in the experimental groups, allowing for the mixing of students of different sexes within tutoring groups.[6]

Within each school-grade stratum, students assigned to a treatment arm were grouped into pairs or triplets based primarily on scheduling constraints, that is, on the basis of their joint availability. The assignment of tutors to these groups followed a similar procedure, with matching determined largely by mutual availability and language spoken by the tutor. This latter restriction was necessary because even within the same school, some classes had as their language of instruction Spanish, while others had Basque, and tutoring sessions had to take place in the same language as the language of instruction.

To allocate tutors to the enhanced socio-emotional training, we randomly selected, from the full pool, the number of tutors needed to serve all students assigned to this treatment. The remaining tutors were allocated to standard tutoring roles. Any leftover tutoring hours for both trained and untrained tutors were then used to provide additional sessions for students outside the experimental evaluation.

[5]The Ministry for Social Inclusion was responsible for the randomized assignment of students to the different treatment arms, with the design and evaluation of the RCT carried out in collaboration with us, as researchers affiliated with CEMFI and J-PAL Europe.

[6]Given the tutoring format (two to three students per tutor) and the stratification constraints, adjustments are necessary to allocate the appropriate number of students to each experimental group within each stratum. In strata that were too small to support all experimental arms, one or more groups were randomly dropped, and students were redistributed proportionally. Within each stratum, assignment followed a random ordering of students and a rotating sequence across groups, with any remaining individuals allocated using the same random sequence at the higher stratum level.

2.7 Timeline

In both waves of the intervention, the following phases were implemented sequentially, as illustrated in Figure 2: First, Navarre's Department for Education contacted eligible schools to inform them about the project and to request their participation. Schools that agreed to participate subsequently informed the families of students enrolled in the specified grades. Families who agreed to participate signed an informed consent.

Once the sample of participating students was finalized, random assignment was carried out to assign students to tutoring groups. Meanwhile, the Department for Education proceeded with the recruitment of tutors. Recruited tutors received preparatory training and were coordinated with the participating schools. Simultaneously, schools administered a baseline test and survey to all participating students.

Following these preparatory steps, the tutorials commenced. The intervention began with an initial face-to-face meeting between tutors and students to facilitate the transition to a digital tutoring environment. Thereafter, the tutoring intervention continued over a period of eight weeks. At the conclusion of the tutoring period, schools administered an endline mathematics test and survey among the participating students.

Recruitment for wave one took place between February and March 2023. Baseline data collection and randomization took place in March, and the intervention started at the end of March and lasted till June. Endline data was collected between end of June and July 2023. For the second wave, recruitment took place from June to July 2023, randomization was done in July and baseline testing took place in September 2023. No communication to families took place during July or August, as schools were closed. Family and student notification is understood to have occurred at the start of the school year in September, around the same time as the baseline test administration for wave 2. We do not have precise documentation of the exact sequencing of notification and test administration, and therefore cannot fully rule out that some students and families learned of their treatment status before completing the baseline assessment. The intervention lasted for eight weeks, from end of September till November 2023. Endline testing and data collection took place between November and December 2023.

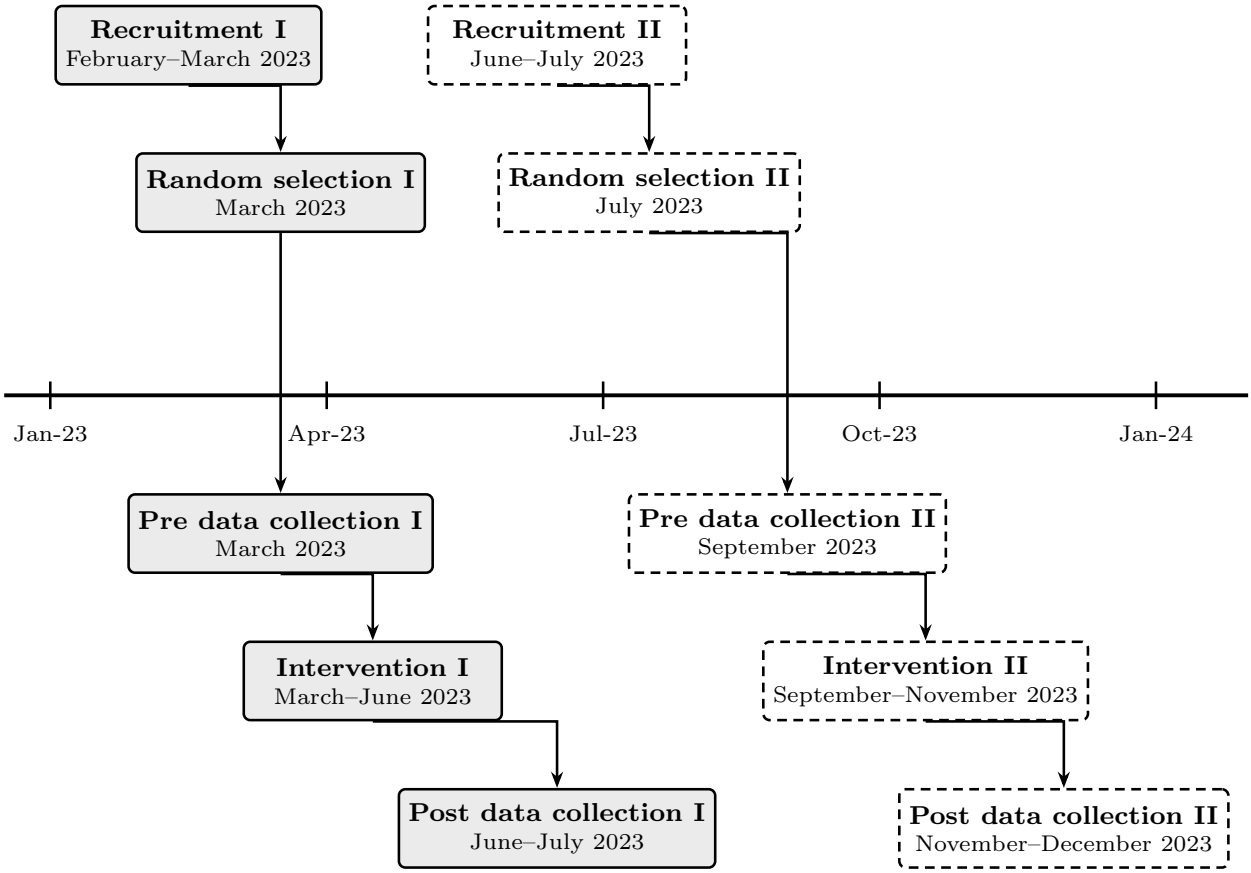
3 Data

We draw on two main sources of data. First, we use administrative records from schools and the Ministry of Education of Navarre. Second, we administered standardized mathematics tests and surveys to students to assess their proficiency in mathematics and assess socio-emotional outcomes.

Administrative records These include information on students' socioeconomic background, vulnerability status (special educational needs census), and academic performance, as reflected in their school grades for different subjects and at different points in time (end-of-quarter and end-of-year).

Mathematics Tests Mathematics assessments were conducted at both baseline and endline. These tests evaluate mathematical competence through problem-solving and calculation tasks. The assessments are tailored to align with students' grade levels to ensure appropriate difficulty and coverage.

Figure 2: Timeline of the intervention



Student Surveys The student surveys collected information on students' perceptions of various socio-educational dimensions and are conducted at two points in time: before the intervention (baseline) and after it concludes (endline). These surveys are administered in classrooms, simultaneously to all participating students. The questionnaire is structured into three main sections: (1) Personal and academic background: This section includes questions on socio-emotional skills, concentration, diligence, academic motivation, and educational aspirations. (2) Relationship with mathematics: This section focuses on students' mathematical abilities, interest in the subject, and how they approach classes and assignments. (3) Well-being and learning environment: This final section addresses students' satisfaction with their social and family life, their perception of school, the importance they assign to education, and other aspects reflecting their relationship with learning. The full list of questions can be found in Online Appendix A.1.

3.1 Descriptive statistics and balancing

Table 1 shows summary statistics and balance checks across treatment and control groups. Most baseline characteristics are well-balanced. However, the treatment group scored 0.2σ higher on the baseline standardized math test ($p = 0.005$), while scoring 0.09σ lower on baseline transcript grades ($p = 0.085$). Although these differences are consistent with random chance rather than systematic imbalance (joint F-test $p = 0.64$),

we address them directly by controlling for the baseline value of each outcome variable in all specifications, which adjusts for pre-existing differences in levels and ensures that our estimates are not confounded by the observed imbalance at baseline.

Primary school students comprised 75% of participants, indicating a strong representation of younger students (average age just over 11 years) in the sample. Consistent with the study's focus on vulnerable populations, 60% of students for whom information is available were listed in the vulnerability census (SESN census), indicating that they had officially recognized special educational needs status. Half of the sample was female, and 67% of students held Spanish nationality. Just over 50% of students were located in rural schools.

3.2 Outcome measures

In the following we define the academic and non-academic outcome measures we consider in our analysis. All outcomes were pre-registered.[7]

Academic outcomes We derive three indicators of academic performance from our data:

1. **Standardized math test score:** A score ranging from 0 to 10 (all correct answers), administered at baseline and endline.

2. **School grades in mathematics:** For Wave 1 (spring 2023), the average grade from the first and second quarters of the 2022-23 academic year serves as the baseline measure, while the final grades of the same year serve as post-intervention observations. For Wave 2 (fall 2023), the third quarter grade of 2022-23 is the baseline, and the first quarter grade of 2023-24 is the post-intervention observation. Quarterly grades range from 0 to 10; final grades range from 0 to 5. We rescale quarterly scores to 0–5 for comparability.

3. **Overall school grades (GPA):** Constructed identically to mathematics grades, using overall performance rather than mathematics alone.

Attitudes toward mathematics We construct two measures of students' self-perceived confidence in their mathematical abilities:

1. **Self-efficacy and math anxiety:** An index constructed from 30 survey items concerning students' relationship with mathematics, perceptions of classes, homework, and exams using Anderson (2008). Each item is rated on a scale of 1 to 5, with higher scores indicating greater self-efficacy and lower anxiety.

2. **Interest in mathematics and language:** Single-item measures assessing students' interest in each subject. Responses range from 1 ("I don't like it") to 5 ("I like it a lot").

[7] The study was preregistered at the AEA RCT registry with entry number AEARCTR-0012727.

Academic aspirations We construct two indicators of future academic intentions from the student questionnaires:

1. **Bachillerato intentions:** A dummy variable indicating whether students believe they will pursue the academic track in upper secondary education.
2. **University aspirations:** A dummy variable indicating whether students expect to attend university after completing school.

Wellbeing at school We construct several indicators of school-related wellbeing from student questionnaires administered at baseline and endline, synthesizing information from multiple questions on each dimension using Anderson (2008)'s method:

1. **School stress:** An index synthesizing information from 11 questions concerning school-related stress.
2. **Well-being and motivation:** A composite indicator based on 14 questions addressing wellbeing and motivation. This indicator is only available for Wave 1.
3. **Grit:** An index based on 10 questions assessing students' ability to confront and sustain their goals, using the Short Grit Scale developed by Duckworth and Quinn (2009).
4. **Life satisfaction:** Constructed from responses to 5 questions regarding students' satisfaction with themselves, friends, family, school, and the environment. Values range from 1 (not at all satisfied) to 4 (very satisfied).

4 Empirical Strategy

To assess the overall impact of the intervention on different student outcomes, we estimate intention-to-treat (ITT) effects using the following specification:

$$y_{ikw} = \alpha_{kw} + \beta_1 \,\text{Tutoring}_i + \gamma \, y_{0i} + \epsilon_{ikw} \tag{1}$$

where y_{ikw} is the outcome for student i in stratum k and wave w, Tutoring_i is a dummy indicating assignment to any tutoring group, y_{0i} is the student's baseline score.[8] We estimate effects on the pooled data for both waves and include strata-wave fixed effects α_{kw}.

The coefficient β_1 captures the average effect of tutoring across both implementation waves. Given that program dropout is extremely low (¡ 1%) and session attendance rates (74%) are comparable to the pilot (80%), the observed 'voltage drop' is unlikely to be attributable to reduced student engagement or attendance challenges but rather to other scaling factors such as tutor selection.

[8] As it is standard in ANCOVA specifications, y_{0i} values are set to zero when the baseline value is missing and we add an indicator variable equal to one when this is the case (McKenzie, 2015).

Does group size matter? To test whether the size of the tutoring group (two vs. three students) influences outcomes, we leverage within-wave variation in group-size assignment. We estimate the following specification, which includes an interaction between the treatment dummy and an indicator for assignment to a group of three students:

$$y_{ikw} = \alpha_{kw} + \beta_1 \,\text{Tutoring}_i + \beta_2 \,(\text{Tutoring}_i \times \text{Group3}_i) + \gamma \, y_{0i} + \epsilon_{ikw} \tag{2}$$

In this equation, Group3_i is a dummy equal to 1 if the student was assigned to a group of three students per tutor. The coefficient β_2 captures the differential effect of being in a group of three relative to a group of two, conditional on being treated.

Does the Socio-emotional Component Improve Tutoring Effectiveness? To assess whether the socio-emotional skills training enhanced the effectiveness of tutoring, we define an indicator variable for exposure to the socio-emotional learning (SEL) component, which equals one for students assigned to tutoring with SEL in Wave 2 and zero otherwise (including all students in Wave 1). Because the SEL module was only implemented in the second wave, we first test whether there are statistically significant differences in effectiveness across waves 1 and 2 using the following estimation equation:

$$y_{ikw} = \alpha_{kw} + \beta_1 \,\text{Tutoring}_i + \beta_2 \,(\text{Tutoring}_i \times \text{Wave2}_i) + \gamma \, y_{0i} + \epsilon_{ikw} \tag{3}$$

To estimate the impact of the additional SEL module on outcomes, we then estimate the following equation using the pooled data from waves 1 and 2:[9]

$$y_{ikw} = \alpha_{kw} + \beta_1 \,\text{Tutoring}_i + \beta_2 \,(\text{Tutoring}_i \times \text{SEL}_i) + \gamma \, y_{0i} + \epsilon_{ikw} \tag{4}$$

SEL_i is a dummy for assignment to the SEL condition. The coefficient β_2 measures the difference between the SEL arm and the pooled overall baseline effect, while β_1 is a regression-weighted average of standard-tutoring effects across waves.

Given the large number of outcome variables we consider and as explained in the previous section, we create summary measures using the inverse covariance–weighted index proposed by Anderson (2008) to mitigate concerns related to multiple hypothesis testing and to increase statistical power.

5 Results

In this section, we will first discuss attrition and potential implications for our results, and then discuss the main findings for academic and socio-emotional outcomes.

[9]We find no statistically significant differences in tutoring effectiveness between waves (we report the interactions from Equation 3 in all of the results tables), justifying the pooled analysis in Equation 4.

5.1 Selective Attrition

Table A.2 reports differences in missingness between treatment and control groups for the main outcomes. As expected given the administrative nature of the data, we observe no differential attrition in grade records. For the standardized mathematics test, overall attrition is low: 10.5% of control group students and 5% of treatment group students have missing scores, a difference of 5.5 percentage points. While statistically significant, this modest differential attrition is likely attributable to the treatment improving school engagement and attendance. In contrast, the socio-emotional outcomes from the voluntary endline survey exhibit higher overall missingness (approximately 55% in the control group) and larger differential attrition (about 17 percentage points lower in the treatment group). Given this differential attrition pattern, particularly for the survey-based outcomes, we report Lee (2009) bounds to account for potential selection bias.

Balancing on Observed Samples Table A.3 reports baseline balance conditional on participation in the endline standardized math test. The response rate was 95% in the treatment group and 90% in the control group. The pattern of baseline differences among test-takers mirrors that of the full sample albeit attenuated: treatment group students scored 0.14σ higher on the baseline standardized math test ($p = 0.069$) and 0.081σ lower on baseline transcript grades ($p = 0.039$). Other baseline covariates are well-balanced. Importantly, we cannot reject the null hypothesis of joint equality of baseline characteristics (F-test p-value = 0.682), suggesting no systematic imbalance.

Table A.4 presents baseline balance conditional on the availability of administrative school records. Attrition was minimal (less than 5% in both groups), and the pattern of baseline characteristics mirrors that of the full sample. The joint F-test (p-value = 0.523) confirms no systematic differences between treatment and control groups.

Table A.5 examines balance among respondents to the endline student survey. Response rates differed substantially between treatment (62%) and control (45%) groups, reflecting the voluntary nature of the survey and differential engagement. Among survey respondents, the pattern of baseline imbalances is consistent with the full sample, though slightly attenuated: treatment students scored 0.16σ higher on the baseline standardized math test ($p = 0.088$) and 0.11σ lower on baseline transcript grades ($p = 0.127$). The joint F-test (p-value = 0.383) confirms no systematic differences in baseline characteristics between treatment and control respondents.

5.2 Program fidelity

Table A.6 captures the attendance of the 971 treatment students to the tutorials, pooled across waves (column 2). Across the two waves, only 5 students who were initially assigned to tutoring never attended any sessions, suggesting an extremely low dropout rate of less than 1%. While nearly no students dropped out entirely, 70% of students assigned to treatment missed at least one session without justified cause. On average, students missed 4.13 out of 16 theoretical sessions they were assigned to attend, implying an attendance

rate of 74% on average, slightly below that in the pilot (80%). Additionally, 5.5% were late for at least one tutoring session (conditional on attending). Columns 3 and 4 show the fidelity indicators separately by wave, showing that attendance improved in the second wave, with a decrease in the percentage of students who missed at least one tutorial (from 76% in wave 1 to 66% in wave 2) and a decrease in the average number of tutorials missed (from 4.6 in wave 1 to 3.8 in wave 2). This improvement likely reflects increased experience and organizational capacity of the implementation team, including more streamlined scheduling processes and better coordination with schools.

5.3 Academic outcomes

Table 2 reports intention-to-treat estimates for the three primary academic outcomes using Eqs. 1-4: standardized math test scores, the math grade from school records, and the overall grade-point average (GPA).

Assignment to tutoring significantly improves two out of three academic outcomes. For the standardized math test, the pooled ITT estimate is 0.2 points ($p < 0.10$, col. 1), equivalent to 0.11σ relative to the control group. For the math grade (0–5 scale), the ITT is 0.21 points ($p < 0.01$, col. 5), or approximately 0.15σ. We do not find a statistically significant effect on overall GPA (0–5 scale), where the estimated ITT effect is 0.04 (or 0.04σ).

These effect sizes correspond to approximately 42% (standardized math test: 0.11 σ vs. 0.26σ) and 31% (math grades: 0.15σ vs. 0.49σ) of the pilot estimates reported in Gortazar et al. (2024), providing a direct measure of the voltage drop associated with the transition from founder-managed pilot to government-implemented scale-up. Meaningful and statistically significant effects survive despite this substantial attenuation across a 3.6-fold expansion, a shift in tutor recruitment, and the absence of founder oversight.

We do not detect meaningful differences by group size: the interaction "Tutoring × 3 students group" is small and not statistically significant for all the academic outcomes (cols. 2, 6, and 10 of Table 2), though our study may be underpowered to detect modest differences. This suggests that substantial cost-savings of around a third could be achieved by increasing group sizes.

Adding the socio-emotional learning (SEL) component does not yield additional gains on academic outcomes. The "Tutoring × SEL" interactions are negative for all three outcomes (cols. 4, 8, and 12 of Table 2)—most notably for the standardized math test—though none reach statistical significance. These estimates indicate that the main treatment effect on academic gains is not driven by the SEL-enhanced variant.

In summary, the core tutoring intervention produces significant gains in math-specific outcomes ranging between 0.11σ and 0.15σ depending on the outcome, with limited evidence of effect heterogeneity across program variations.

5.4 Socio-emotional outcomes and aspirations

We now examine impacts on attitudes toward mathematics, school-related stress and wellbeing, grit, and educational aspirations, using Eqs. 1-4 and the composite indices described in Section 3.2.

Tutoring increased the composite index capturing higher self-efficacy and lower anxiety in math by 0.22σ ($p < 0.05$) (see Table 3, column 1). There are no significant heterogeneities with respect to group size (column 2). However, the socio-emotional treatment arm amplified the gains from tutoring sessions: the "Tutoring × SEL" interaction is 0.22σ ($p < 0.10$), indicating an additional improvement on top of standard tutoring.

We do not find evidence of significant effects on the measure of interest in mathematics or on the standardized school stress index (Table 3 columns 5-12). The exception is the socio-emotional treatment arm, where we find that interest in math increases significantly (about 0.3σ), while students who receive the baseline treatment do not experience any gains. For broader wellbeing (only measured during wave 1) and grit (see Table 4), average impacts are close to zero and imprecise across specifications, and we do not observe significant heterogeneity by group size or SEL assignment.

Turning to future academic aspirations (see Table 5), we study self-reported intentions to pursue the academic upper-secondary track (Bachillerato) and to attend university. In the pooled sample, tutoring has small and statistically insignificant effects on both outcomes. However, the SEL interaction is negative and statistically significant for university aspirations, where students assigned to this treatment arm report a 15-percentage point lower likelihood of planning to attend university than those in the baseline treatment ($p < 0.01$).

Finally, we look at whether life satisfaction is affected by the intervention. Table 6 shows that overall, the intervention did not improve life satisfaction, but had a positive and significant impact of 0.16σ for the SEL treatment arm. We further find an increased interest in language. Although across most specifications the impact is imprecisely estimated, coefficients are large and positive and are significant for the treatment arm with two students per tutor ($+0.17\sigma$) and large, positive and significant for the SEL treatment arm ($+0.26\sigma$).

In sum, while the standard tutoring intervention shows limited impacts on socio-emotional outcomes beyond math self-efficacy, the SEL-enhanced treatment consistently demonstrates broader benefits across affective domains—improving math self-efficacy and anxiety, interest in mathematics, life satisfaction, and interest in language. This pattern suggests that explicit socio-emotional skill development may be necessary to generate meaningful non-academic gains, even if it does not translate into additional academic achievement in the short term (Jackson, 2018; Algan and Huillery, 2025). However, the finding that SEL students report lower university aspirations suggests potential unintended consequences that warrant further investigation.

5.5 Robustness to differential attrition

Table A.7 presents Lee bounds to assess the robustness of our main treatment effects to differential attrition. For math grades from administrative records—where attrition was minimal—the bounds are tight and identical to the point estimate (0.210, all $p < 0.01$), confirming the result is not sensitive to selection. For the standardized math test, the lower bound (0.075) is no longer statistically significant, while the upper bound (0.304) remains highly significant, suggesting moderate sensitivity to assumptions about how treatment affected attrition. The bounds remain positive throughout, consistent with a beneficial effect, though the magnitude is less precisely identified. For socio-emotional outcomes, where attrition was substantial, the bounds are considerably wider, ranging from 0 to 0.649 (highly significant). This wide range reflects the large differential response rates on the voluntary survey and indicates that estimates for these outcomes should be interpreted with greater caution. Overall, the core finding of positive academic impacts is robust for transcript grades and plausible (though less certain) for test scores, while socio-emotional effects are more sensitive to selection assumptions.

5.6 Heterogeneous Treatment Effects

Table 7 explores treatment effect heterogeneity across student characteristics. Overall, we find limited systematic variation in program impacts, though several patterns emerge.

Gender and nationality We detect no significant differences in treatment effects by gender across any outcome. For students of Spanish nationality, the interaction is positive but insignificant for the standardized math test, while non-Spanish students show near-zero effects. Spanish students also report a significant decline in interest in mathematics relative to non-Spanish students (for whom the effect is positive but insignificant), suggesting divergent experiences within the program.

Urban vs. rural context Treatment effects on the standardized math test appear concentrated among urban students (17% of the sample), with a large positive interaction, while rural students show minimal gains. This heterogeneity does not extend to transcript grades. For socio-emotional outcomes, the pattern reverses: the negative interaction for self-efficacy and anxiety suggests near-zero effects for urban students, while rural students benefit.

School level We find that treatment effects on academic outcomes are concentrated among primary school students. For the standardized math test, the combined effect for secondary students is close to zero and substantially below the primary school estimate, though the interaction is imprecisely estimated and does not reach statistical significance. For math grades, the pattern is similar: the point estimate for secondary students is smaller, though again the difference is not statistically significant. In contrast, secondary students show larger gains in math self-efficacy and anxiety, suggesting that older students may respond to the intervention more through affective than academic channels—a pattern worth investigating in future work but which should be interpreted cautiously given limited statistical power.

Baseline achievement For the standardized math test, treatment effects appear concentrated among students who scored above the median at baseline: the combined effect for this group is 0.19σ (p ¡ 0.05), while students below the median show a near-zero and insignificant effect. The interaction term itself is large (0.308) but imprecisely estimated and does not reach statistical significance, so this pattern should be interpreted with caution. For transcript grades, a similar gradient is evident, with larger effects for students scoring in the upper half of the baseline achievement distribution.

In sum, while we find limited systematic heterogeneity overall, several patterns emerge consistently across outcomes. Academic gains from tutoring appear more concentrated among primary school students and those with higher baseline achievement, while secondary students and lower-performing peers tend to show stronger affective responses. Urban students drive the standardized test gains, whereas rural students benefit more on socio-emotional outcomes. Effects do not differ meaningfully by gender, and nationality differences are limited to interest in mathematics. Importantly, most interaction terms are imprecisely estimated and fall short of statistical significance, so these patterns should be read as suggestive rather than conclusive. They point to potentially important heterogeneity in how different student groups respond to tutoring across academic and affective domains, but replication with adequately powered studies is needed before drawing firm conclusions.

Tutor characteristics Table 8 examines heterogeneous effects by tutor characteristics: age, gender, and prior experience (as tutor, teacher, or volunteer). We find limited variation in treatment effects across most tutor attributes, with a few exceptions for academic outcomes.

For the standardized math test, positive effects are concentrated among tutors above the median age, female tutors and those with prior tutoring experience, while general teaching experience or volunteer experience do not significantly moderate impacts.

In contrast, we find no heterogeneity by tutor age, gender or prior tutor experience for math transcript grades. Similarly, socio-emotional outcomes—attitudes toward math and school stress—are not significantly affected by any measured tutor characteristics.

Overall, tutor background appears to matter primarily for performance on the standardized assessment, transcript grades in math and the self-efficacy and anxiety index, with older, more experienced tutors (those with prior teaching, tutoring or volunteer experience) and female tutors showing stronger impacts— though these patterns should not be interpreted as causal given that tutors were not randomly assigned to students.

6 Conclusion

This study examines whether an effective online mathematics tutoring program can survive the transition from a carefully controlled pilot to a 3.6-fold scaled-up government intervention, offering a demanding test of whether evidence-based programs retain their impact under realistic implementation conditions.

Our findings demonstrate that meaningful impacts persist at scale, though with substantial attenuation.

Assignment to tutoring increased standardized mathematics test scores by 0.11σ and transcript grades by 0.15σ, representing approximately 30-42 percent of the original pilot effects. Students also reported significantly higher math self-efficacy and lower anxiety (0.22σ). These results confirm the "voltage drop" documented across educational interventions, but show that a core of effectiveness survives even without founder involvement, with a different workforce, and under full government management.

Experimental variation in group size further suggests that expanding student-tutor ratios from 2:1 to 3:1 does not meaningfully attenuate the observed gains, though we lack statistical power to rule out modest differences in effectiveness. If this null result is taken at face value, it would imply a straightforward path to reducing per-student costs by approximately one-third—a finding that warrants replication with adequately powered studies before informing policy. The enhanced socio-emotional training module, while not improving academic outcomes beyond the standard program, generated meaningful additional benefits across affective domains—including math interest and life satisfaction—suggesting it may be worth incorporating when broader wellbeing goals are part of the policy objective.

The online provision of the program together with the staffing model evaluated here addresses perhaps the most binding constraint on tutoring scale-up: tutor supply. By leveraging interim teachers—qualified educators already integrated into the government's teacher registry, who face employment gaps between permanent assignments—this approach eliminates recruitment bottlenecks, political economy challenges, and reduces marginal costs, while allowing to bring tutoring at scale to remote rural areas. That these readily available educators generate meaningful learning gains, even without intensive screening, suggests that many education systems might possess untapped capacity for intensive intervention.

Several limitations warrant acknowledgment. Our outcome measurement occurred immediately post-intervention; longer-run persistence remains unknown. Differential attrition on survey measures, though addressed through bounding exercises, introduces uncertainty around socio-emotional estimates. The finding that SEL-trained tutors' students report lower university aspirations requires further investigation. And generalizability to other contexts remains an open question.

However, our findings carry clear policy implications. First, policymakers should anticipate voltage drop when scaling evidence-based interventions but recognize that meaningful effects can persist—the relevant question is whether attenuated effects justify costs under realistic implementation. Second, existing civil service structures can be repurposed for intensive intervention avoiding the administrative frictions and fixed costs that come with external recruitment, particularly in contexts with large pools of qualified but underemployed educators. Third, online delivery eliminates physical infrastructure constraints, enabling flexible scheduling and broader reach.

Future research should examine longer-run effects, cost-effectiveness relative to alternative staffing models, and the mechanisms through which tutoring operates. In sum, this study demonstrates that online tutoring can generate meaningful benefits when implemented at scale by government authorities using existing staff. For policymakers seeking scalable strategies to address learning gaps, mobilizing underutilized teaching

capacity represents a viable path forward.

References

Agostinelli, Francesco, Ciro Avitabile, and Matteo Bobba, "Enhancing Human Capital in Children: A Case Study on Scaling," *Journal of Political Economy*, 2025, *133* (2), 455–491.

Ajzenman, Nicolás, Gregory Elacqua, Analía Jaimovich, and Graciela Pérez-Núñez, "Humans versus Chatbots: Scaling-Up Behavioral Interventions to Reduce Teacher Shortages," *Journal of Political Economy Microeconomics.*, 2024. Forthcoming.

Al-Ubaydli, Omar, John A. List, and Dana L. Suskind, "What Can We Learn from Experiments? Understanding the Threats to the Scalability of Experimental Results," *American Economic Review*, May 2017, *107* (5), 282–86.

Algan, Yann and Elise Huillery, "Socio-Emotional Skills and the Future of Education," *Annual Review of Economics*, 2025, *17*.

Anderson, Michael L., "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American Statistical Association*, 2008, *103* (484), 1481–1495.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton, "From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application," *Journal of Economic Perspectives*, 2017, *31* (4), 73–102.

_ , _ , _ , _ , _ , _ , _ , and _ , "From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application," *Journal of Economic Perspectives*, November 2017, *31* (4), 73–102.

Bettinger, Eric P., Lindsay Fox, Susanna Loeb, and Eric S. Taylor, "Virtual Classrooms: How Online College Courses Affect Student Success," *American Economic Review*, September 2017, *107* (9), 2855–75.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Justin Sandefur et al., "Experimental evidence on scaling up education reforms in Kenya," *Journal of Public Economics*, 2018, *168*, 1–20.

Brown, Lindsay E., Ha Yeon Kim, Carly Tubbs Dolan, Autumn Brown, Jennifer Sklar, and J. Lawrence Aber, "Remedial Programming and Skill-Targeted SEL in Low-Income and Crisis-Affected Contexts: Experimental Evidence From Niger," *Journal of Research on Educational Effectiveness*, 2023, *16* (4), 583–614.

Carlana, Michela and Eliana La Ferrara, "Apart but Connected: Online Tutoring, Cognitive Outcomes, and Soft Skills," *American Economic Review*, October 2025, *115* (10), 3487–3513.

Duckworth, A. L. and P. D. Quinn, "Development and validation of the Short Grit Scale (GRIT–S)," *Journal of Personality Assessment*, 2009, *91* (2), 166—-174.

Durlak, Joseph A, Roger P Weissberg, Allison B Dymnicki, Rebecca D Taylor, and Kriston B Schellinger, "The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions," *Child Development*, 2011, *82* (1), 405–432.

Escueta, Maya, Andre Joshua Nickow, Philip Oreopoulos, and Vincent Quan, "Upgrading Education with Technology: Insights from Experimental Research," *Journal of Economic Literature*, December 2020, *58* (4), 897–996.

European Committee of the Regions, "Spain - Education," `https://portal.cor.europa.eu/divisionpowers/Pages/spain-edu.aspx` 2023. Accessed: 2025-04-22.

Gobierno de Navarra, Departamento de Educación, "Estadística de Datos Básicos," `https://www.educacion.navarra.es/web/dpto/estadisticas/estadistica-de-datos-basicos` 2024. Accessed: 2025-04-22.

_ , "Programa Laguntza para el bienestar emocional," `https://convivencia.educacion.navarra.es/programa-laguntza` 2025. Accessed: 2025-06-12.

Gortazar, Lucas, Claudia Hupkau, and Antonio Roldán-Monés, "Online tutoring works: Experimental evidence from a program with vulnerable children," *Journal of Public Economics*, 2024, *232*, 105082.

Guryan, Jonathan, Jens Ludwig, Monica P. Bhatt, Philip J. Cook, Jonathan M. V. Davis, Kenneth Dodge, George Farkas, Jr. Fryer Roland G., Susan Mayer, Harold Pollack, Laurence Steinberg, and Greg Stoddard, "Not Too Late: Improving Academic Outcomes among Adolescents," *American Economic Review*, March 2023, *113* (3), 738–65.

Hardt, David, Markus Nagler, and Johannes Rincke, "Tutoring in (Online) Higher Education: Experimental Evidence," *Economics of Education Review*, 2023, *92*, 102350.

Jackson, C Kirabo, "What do test scores miss? The importance of teacher effects on non–test score outcomes," *Journal of Political Economy*, 2018, *126* (5), 2072–2107.

Justman, Moshe, "Randomized controlled trials informing public policy: Lessons from Project STAR and class size reduction," *European Journal of Political Economy*, 2018, *54*, 167–174.

Kraft, Matthew A. and Grace T. Falken, "A Blueprint for Scaling Tutoring and Mentoring Across Public Schools," *AERA Open*, 2021, *7*, 23328584211042858.

_ **and Virginia S. Lovison**, "The Effect of Student–Tutor Ratios: Experimental Evidence From a Pilot Online Math Tutoring Program," *Educational Evaluation and Policy Analysis*, 2025. OnlineFirst.

Kraft, Matthew A., Beth E. Schueler, and Grace Falken, "What Impacts Should We Expect from Tutoring at Scale? Exploring Meta-Analytic Generalizability," August 2024, (1031).

_ **, Danielle Sanderson Edwards, and Marisa Cannata**, "The Scaling Dynamics and Causal Effects of a District-Operated Tutoring Program," August 2024, (1030).

Kraft, Matthew A., John A. List, Jeffrey A. Livingston, and Sally Sadoff, "Online Tutoring by College Volunteers: Experimental Evidence from a Pilot Program," *American Economic Association*, 2022, *112*, 614–618.

Lee, David S., "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *The Review of Economic Studies*, 07 2009, *76* (3), 1071–1102.

List, John A., *The Voltage Effect: How to Make Good Ideas Great and Great Ideas Scale*, New York, NY: Currency / Penguin Random House, 2022.

McKenzie, David, "Another reason to prefer ANCOVA: dealing with changes in measurement between baseline and follow-up," Development Impact (World Bank Blogs) June 2015.

Muralidharan, Karthik and Paul Niehaus, "Experimentation at Scale," *Journal of Economic Perspectives*, November 2017, *31* (4), 103–24.

Nickow, Andre, Philip Oreopoulos, and Vincent Quan, "The Promise of Tutoring for PreK–12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence," *American Educational Research Journal*, 2024, *61* (1), 74–107.

Noticias de Navarra, "El programa Laguntza se reenfoca hacia la educación socioemocional y el buen trato en las aulas," 2023. Accessed: 2025-06-12.

OECD, *Working and Learning Together: Rethinking Human Resource Policies for Schools* OECD Reviews of School Resources, Paris: OECD Publishing, 2019.

Vivalt, Eva, "How Much Can We Generalize From Impact Evaluations?," *Journal of the European Economic Association*, 09 2020, *18* (6), 3045–3089.

Tables

Table 1: Students characteristics at baseline

	(1) Obs.	(2) All	(3) Control	(4) Treatment	(5) P-value	(6) Std. diff.
Students characteristics						
Female	1344	0.513	0.515	0.513	0.377	-0.004
Age	1344	11.458	11.496	11.444	0.681	-0.035
5 EP	1344	0.435	0.429	0.437		0.016
6 EP	1344	0.319	0.340	0.311		-0.063
1 ESO	1344	0.141	0.131	0.144		0.037
2 ESO	1344	0.106	0.099	0.108		0.029
Spanish nationality	1344	0.670	0.670	0.670	0.701	0.000
SESN census	996	0.597	0.598	0.597	0.958	-0.003
School in an urban area	1344	0.161	0.169	0.158		-0.031
School in a semin-dense area	1344	0.332	0.343	0.327		-0.033
School in a rural area	1344	0.507	0.488	0.515		0.054
Academic outcomes						
Maths - standardized test	1344	2.851	2.602	2.947	0.005	0.195
Maths grade - school transcripts	1344	2.295	2.389	2.258	0.085	-0.089
Grade average - school transcripts	1344	2.841	2.869	2.830	0.321	-0.034
Socio-emotional outcomes and aspirations						
Self-efficacy and maths anxiety index	1344	1.386	1.210	1.454	0.268	0.120
Interest in maths	969	3.437	3.466	3.426	0.707	-0.032
School stress index	1344	1.158	0.932	1.246	0.004	0.195
School well-being and motivation index	545	0.310	0.312	0.309	0.715	-0.003
Grit index	1344	1.174	0.957	1.257	0.009	0.187
Aspiration to study *Bachillerato*	1344	0.218	0.241	0.209	0.724	-0.077
Aspiration to attend university	1344	0.317	0.340	0.308	0.694	-0.070
Life satisfaction index	1344	1.296	1.102	1.371	0.032	0.155
Interest in language (Spanish/Basque)	962	3.739	3.736	3.740	0.930	0.004
Joint test (p-value)					0.638	

Notes: Column 1 reports the number of observations. Columns 2–4 show the mean for the full sample, the control group, and the treatment group, respectively. Column 5 presents p-values from regressions testing whether the treatment–control mean difference is zero, controlling for stratification and wave fixed effects and using robust standard errors. The last row reports the p-value from a joint test of overall significance from the same specification. Column 6 shows the standardized difference in means, calculated as the difference in means divided by the pooled standard deviation. For a variable X, let $\bar{X}_1$ and $\bar{X}_0$ denote the sample means in groups 1 and 0, and let s_1 and s_0 denote the corresponding sample standard deviations. The pooled standard deviation is $s_{\text{pooled}} = \sqrt{\frac{s_0^2+s_1^2}{2}}$. The standardized difference is then $\text{StdDiff} = \frac{\bar{X}_1-\bar{X}_0}{s_{\text{pooled}}}$. *** p<0.01, ** p<0.05, * p<0.1.

Table 2: Impact of Tutoring on Academic Outcomes

	Maths standardized test				School transcripts: Maths grade				School transcripts: Grade average			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Tutoring	0.200*	0.188*	0.155	0.224**	0.210***	0.199***	0.126**	0.219***	0.038	0.032	0.002	0.046*
	(0.104)	(0.112)	(0.131)	(0.106)	(0.051)	(0.057)	(0.052)	(0.052)	(0.027)	(0.029)	(0.037)	(0.027)
Tutoring × 3 students group		0.029				0.027				0.015		
		(0.101)				(0.058)				(0.029)		
Tutoring × Wave 2			0.084				0.158				0.069	
			(0.205)				(0.099)				(0.053)	
Tutoring × SEL				-0.144				-0.051				-0.043
				(0.130)				(0.085)				(0.042)
Observations	1256	1256	1256	1256	1302	1302	1302	1302	1302	1302	1302	1302
R^2	0.534	0.534	0.534	0.535	0.756	0.756	0.757	0.756	0.855	0.855	0.855	0.855
Control mean dep. var.	3.180	3.180	3.180	3.180	2.417	2.417	2.417	2.417	2.991	2.991	2.991	2.991
Control SD dep. var.	1.868	1.868	1.868	1.868	1.430	1.430	1.430	1.430	0.931	0.931	0.931	0.931

Notes: The table shows the main results from estimation Eqs. 1-4. Columns (1)-(4) present effects on standardized math test scores, columns (5)-(8) on math grades (from school transcripts), and columns (9)-(12) on overall grade point average (from school transcripts). Column (1), (5), and (9) show pooled ITT estimates. Columns (2), (6), and (10) include interactions with group size (3 students). Columns (3), (7), and (11) include interactions with Wave 2 timing (Spring). Columns (4), (8), and (12) include interactions with the SEL treatment arm. All specifications control for baseline values of the outcome variable. Robust standard errors are presented in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: Impact of Tutoring on Attitudes and School Stress

	Self-efficacy & anxiety index				Interest in math				School stress index			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Tutoring	0.271***	0.244**	0.224	0.231**	-0.079	-0.070	-0.068	-0.114	-0.111	-0.104	-0.079	-0.101
	(0.102)	(0.115)	(0.137)	(0.104)	(0.111)	(0.115)	(0.152)	(0.114)	(0.089)	(0.097)	(0.125)	(0.092)
Tutoring × 3 students group		0.064				-0.021				-0.018		
		(0.115)				(0.102)				(0.091)		
Tutoring × Wave 2			0.105				-0.030				-0.067	
			(0.205)				(0.214)				(0.177)	
Tutoring × SEL				0.278*				0.314**				-0.063
				(0.149)				(0.141)				(0.113)
Observations	616	616	616	616	618	618	618	618	723	723	723	723
R^2	0.577	0.577	0.577	0.580	0.357	0.357	0.357	0.363	0.434	0.434	0.434	0.434
Control mean dep. var.	4.802	4.802	4.802	4.802	3.806	3.806	3.806	3.806	2.702	2.702	2.702	2.702
Control SD dep. var.	1.243	1.243	1.243	1.243	1.076	1.076	1.076	1.076	1.004	1.004	1.004	1.004

Notes: The table shows the main results from estimation Eqs. 1-4 for socio-emotional outcomes. Columns (1)-(4) present effects on the self-efficacy and anxiety index (higher values indicate greater self-efficacy and lower anxiety), columns (5)-(8) on interest in mathematics, and columns (9)-(12) on the school stress index. Columns (1), (5), and (9) show pooled ITT estimates. Columns (2), (6), and (10) include interactions with group size (3 students). Columns (3), (7), and (11) include interactions with Wave 2 timing. Columns (4), (8), and (12) include interactions with the SEL treatment arm. All specifications control for baseline values of the outcome variable. Robust standard errors are presented in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: Impact of Tutoring on Wellbeing and Grit

	Wellbeing		Grit			
	(1)	(2)	(3)	(4)	(5)	(6)
Tutoring	0.081	0.003	0.001	0.007	0.027	-0.027
	(0.124)	(0.136)	(0.091)	(0.098)	(0.122)	(0.093)
Tutoring × 3 students group		0.151		-0.015		
		(0.128)		(0.083)		
Tutoring × Wave 2					-0.056	
					(0.183)	
Tutoring × SEL						0.173
						(0.110)
Observations	320	320	735	735	735	735
R^2	0.461	0.464	0.633	0.633	0.633	0.635
Control mean dep. var.	2.933	2.933	3.194	3.194	3.194	3.194
Control SD dep. var.	1.000	1.000	1.190	1.190	1.190	1.190

Notes: The table shows the main results from estimation equations (1)-(4) for wellbeing and grit. Columns (1)-(2) present effects on wellbeing, with column (1) showing the pooled ITT estimate and column (2) including interactions with group size (3 students). Wellbeing was only measured in Wave 1, so Wave 2 and SEL interactions are not included for this outcome. Columns (3)-(6) present effects on grit. Column (3) shows the pooled ITT estimate, column (4) includes interactions with group size, column (5) includes interactions with Wave 2 timing, and column (6) includes interactions with the SEL treatment arm. All specifications control for baseline values of the outcome variable. Robust standard errors are presented in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: Impact of Tutoring on Academic Aspirations

	Study *Bachillerato*				Attend university			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Tutoring	0.012	0.028	-0.019	0.020	0.032	0.026	0.053	0.055
	(0.045)	(0.048)	(0.061)	(0.046)	(0.041)	(0.044)	(0.059)	(0.042)
Tutoring × 3 students group		-0.037				0.013		
		(0.043)				(0.038)		
Tutoring × Wave 2			0.066				-0.045	
			(0.091)				(0.082)	
Tutoring × SEL				-0.052				-0.146***
				(0.064)				(0.055)
Observations	768	768	768	768	766	766	766	766
R^2	0.333	0.333	0.333	0.333	0.338	0.338	0.338	0.347
Control mean dep. var.	0.588	0.588	0.588	0.588	0.706	0.706	0.706	0.706
Control SD dep. var.	0.494	0.494	0.494	0.494	0.457	0.457	0.457	0.457

Notes: The table shows the main results from estimation equations (1)-(4) for academic aspirations. Columns (1)-(4) present effects on students' intentions to study *Bachillerato* (upper secondary school track that provides entry to university), and columns (5)-(8) on intentions to attend university. Columns (1) and (5) show pooled ITT estimates. Columns (2) and (6) include interactions with group size (3 students). Columns (3) and (7) include interactions with Wave 2 timing. Columns (4) and (8) include interactions with the SEL treatment arm. All specifications control for baseline values of the outcome variable. Robust standard errors are presented in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 6: Impact of Tutoring on Satisfaction and Interest in Language

	Satisfaction				Interest in language			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Tutoring	0.057	0.084	0.087	0.022	0.144	0.215**	0.169	0.107
	(0.099)	(0.105)	(0.133)	(0.101)	(0.097)	(0.100)	(0.120)	(0.099)
Tutoring × 3 students group		-0.066				-0.165*		
		(0.088)				(0.095)		
Tutoring × Wave 2			-0.064				-0.066	
			(0.200)				(0.204)	
Tutoring × SEL				0.226**				0.324***
				(0.100)				(0.120)
Observations	745	745	745	745	620	620	620	620
R^2	0.670	0.670	0.670	0.672	0.618	0.620	0.618	0.623
Control mean dep. var.	3.139	3.139	3.139	3.139	3.597	3.597	3.597	3.597
Control SD dep. var.	1.441	1.441	1.441	1.441	1.232	1.232	1.232	1.232

Notes: The table shows the main results from estimation equations (1)-(4) for life satisfaction and interest in language. Columns (1)-(4) present effects on life satisfaction, and columns (5)-(8) on interest in language. Columns (1) and (5) show pooled ITT estimates. Columns (2) and (6) include interactions with group size (3 students). Columns (3) and (7) include interactions with Wave 2 timing. Columns (4) and (8) include interactions with SEL treatment arm. Life satisfaction is constructed from responses to 5 questions regarding students' satisfaction with themselves, friends, family, school, and the environment, ranging from 1 (not at all satisfied) to 4 (very satisfied); higher values indicate greater satisfaction. Interest in language assesses students' level of liking for the subject, ranging from 1 (I don't like it) to 5 (I like it a lot); higher values indicate higher interest. All specifications control for baseline values of the outcome variable. Robust standard errors are presented in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 7: Heterogeneous treatment effects

	Academic Outcomes (Table 2)			Attitudes and School Stress (Table 3)		
	Maths stndrd. test	Maths grade	Grade average	Self-efficacy & anxiety index	Interest in math	School stress index
	(1)	(2)	(3)	(4)	(5)	(6)
Tutoring	0.141	0.195***	0.040	0.307**	0.022	-0.234*
	(0.147)	(0.072)	(0.040)	(0.152)	(0.165)	(0.128)
Tutoring × Female	0.120	0.027	-0.003	-0.068	-0.214	0.237
	(0.209)	(0.102)	(0.053)	(0.205)	(0.221)	(0.178)
Observations	1256	1302	1302	616	618	723
Tutoring+Tutoring× Female	0.26*	0.22***	0.04	0.24*	-0.19	0.00
	(0.15)	(0.07)	(0.04)	(0.14)	(0.15)	(0.12)
Tutoring	0.021	0.262***	0.079	0.126	0.255	-0.154
	(0.183)	(0.096)	(0.049)	(0.183)	(0.237)	(0.153)
Tutoring × Spanish	0.272	-0.077	-0.060	0.224	-0.485*	0.061
	(0.229)	(0.117)	(0.060)	(0.228)	(0.269)	(0.185)
Observations	1256	1302	1302	616	618	723
Tutoring+Tutoring× Spanish	0.29**	0.18***	0.02	0.35***	-0.23*	-0.09
	(0.13)	(0.06)	(0.03)	(0.13)	(0.12)	(0.11)
Tutoring	0.070	0.225***	0.044	0.354***	0.012	-0.151
	(0.114)	(0.055)	(0.029)	(0.117)	(0.135)	(0.102)
Tutoring × Urban	0.789***	-0.084	-0.034	-0.331	-0.354	0.163
	(0.260)	(0.148)	(0.071)	(0.238)	(0.226)	(0.209)
Observations	1256	1302	1302	616	618	723
Tutoring+Tutoring× Urban	0.86***	0.14	0.01	0.02	-0.34*	0.01
	(0.23)	(0.14)	(0.06)	(0.21)	(0.18)	(0.18)
Tutoring	0.249**	0.229***	0.034	0.191*	-0.043	-0.143
	(0.121)	(0.061)	(0.030)	(0.111)	(0.118)	(0.095)
Tutoring × Secondary	-0.197	-0.077	0.018	0.650**	-0.314	0.270
	(0.231)	(0.108)	(0.065)	(0.253)	(0.329)	(0.274)
Observations	1256	1302	1302	616	618	723
Tutoring+Tutoring× Secondary	0.05	0.15*	0.05	0.84***	-0.36	0.13
	(0.20)	(0.09)	(0.06)	(0.23)	(0.31)	(0.26)
Tutoring	0.043	0.157**	-0.013	0.251	-0.154	-0.129
	(0.149)	(0.076)	(0.039)	(0.161)	(0.176)	(0.124)
Tutoring × Maths test score $\geq 50p$	0.308	0.081	0.086	0.007	0.151	0.060
	(0.210)	(0.106)	(0.053)	(0.210)	(0.225)	(0.178)
Observations	1256	1302	1302	616	618	723
Tutoring+Tutoring× Maths test score $\geq 50p$	0.35**	0.24***	0.07**	0.26**	-0.00	-0.07
	(0.15)	(0.07)	(0.04)	(0.13)	(0.14)	(0.13)
Control mean dep. var.	3.180	2.417	2.991	4.802	3.806	2.702
Control SD dep. var.	1.868	1.430	0.931	1.243	1.076	1.004

Notes: The table presents heterogeneous treatment effects for main outcomes from Eq. 1 plus interactions of student characteristics with the main effect. All specifications control for baseline values of the outcome variable. Robust standard errors are presented in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 8: Heterogeneous treatment effects by tutor characteristics

	Academic Outcomes (Table 2)			Attitudes and School Stress (Table 3)		
	Maths stndrd. test	Maths grade	Grade average	Self-efficacy & anxiety index	Interest in math	School stress index
	(1)	(2)	(3)	(4)	(5)	(6)
Tutoring	0.052	0.206***	0.049	0.310***	-0.050	-0.149
	(0.118)	(0.060)	(0.032)	(0.120)	(0.118)	(0.102)
Tutoring × Age above median	0.290**	0.008	-0.021	-0.108	-0.081	0.106
	(0.124)	(0.071)	(0.036)	(0.150)	(0.139)	(0.125)
Observations	1256	1302	1302	616	618	723
Tutoring+Tutoring× Age above median	0.34***	0.21***	0.03	0.20	-0.13	-0.04
	(0.12)	(0.06)	(0.03)	(0.13)	(0.15)	(0.12)
Tutoring	0.008	0.183**	0.053	0.317	-0.248	-0.100
	(0.160)	(0.086)	(0.045)	(0.204)	(0.181)	(0.156)
Tutoring × Female tutor	0.253	0.028	-0.010	-0.013	0.208	-0.052
	(0.163)	(0.090)	(0.047)	(0.210)	(0.186)	(0.145)
Observations	1127	1167	1167	554	558	653
Tutoring+Tutoring× Female tutor	0.26**	0.21***	0.04	0.30***	-0.04	-0.15
	(0.11)	(0.06)	(0.03)	(0.11)	(0.12)	(0.09)
Tutoring	0.036	0.191***	0.020	0.294*	-0.053	-0.112
	(0.130)	(0.066)	(0.039)	(0.163)	(0.155)	(0.135)
Tutoring × Experience: tutor	0.271*	0.027	0.045	0.014	-0.039	-0.034
	(0.142)	(0.079)	(0.044)	(0.188)	(0.158)	(0.140)
Observations	1158	1202	1202	567	569	666
Tutoring+Tutoring× Experience: tutor	0.31***	0.22***	0.06**	0.31**	-0.09	-0.15
	(0.12)	(0.06)	(0.03)	(0.13)	(0.13)	(0.10)
Tutoring	0.212*	0.151**	0.028	0.382***	-0.066	-0.207*
	(0.123)	(0.065)	(0.034)	(0.130)	(0.149)	(0.110)
Tutoring × Experience: voluntary	-0.019	0.109	0.038	-0.124	-0.030	0.142
	(0.133)	(0.077)	(0.040)	(0.163)	(0.158)	(0.120)
Observations	1146	1188	1188	558	561	659
Tutoring+Tutoring× Experience: voluntary	0.19	0.26***	0.07*	0.26*	-0.10	-0.06
	(0.12)	(0.06)	(0.03)	(0.14)	(0.13)	(0.11)
Tutoring	0.263	0.276*	0.025	0.665**	-0.116	0.250
	(0.242)	(0.151)	(0.083)	(0.333)	(0.624)	(0.388)
Tutoring × Experience: teacher	-0.073	-0.074	0.024	-0.370	0.039	-0.394
	(0.244)	(0.155)	(0.085)	(0.335)	(0.625)	(0.385)
Observations	1158	1202	1202	567	569	666
Tutoring+Tutoring× Experience: teacher	0.19*	0.20***	0.05*	0.30***	-0.08	-0.14
	(0.11)	(0.05)	(0.03)	(0.11)	(0.12)	(0.09)
Control mean dep. var.	3.180	2.417	2.991	4.802	3.806	2.702
Control SD dep. var.	1.868	1.430	0.931	1.243	1.076	1.004

Online Appendix

A.1 Student survey

PRE-TEST AND POST-TEST QUESTIONNAIRE FOR STUDENTS (EXTENDED VERSION)

General instructions: Please answer honestly. There are no right or wrong answers.

A- Socio-emotional skills (Grit)

Below are a series of statements that may or may not apply to you. There are no right or wrong answers, so please answer honestly, keeping in mind how you compare to most people.

P1- Choose one of the following options: "Very much like me", "Mostly like me", "Somewhat like me", "Not much like me" and "Not like me at all":

		Very much like me	Mostly like me	Somewhat like me	Not much like me	Not like me at all
1.	New ideas and projects sometimes distract me from previous ones.	❑	❑	❑	❑	❑
2.	Setbacks do not discourage me. I do not give up easily.	❑	❑	❑	❑	❑
3.	I often set a goal, but then choose to pursue a different one.	❑	❑	❑	❑	❑
4.	I am a hard worker.	❑	❑	❑	❑	❑
5.	I find it difficult to stay focused on projects that take more than a few months to complete.	❑	❑	❑	❑	❑
6.	I finish whatever I start.	❑	❑	❑	❑	❑
7.	My interests change from year to year.	❑	❑	❑	❑	❑
8.	I am diligent. I never give up.	❑	❑	❑	❑	❑
9.	I have been obsessed with a certain idea or project for a short time, but then I lost interest.	❑	❑	❑	❑	❑
10.	I have overcome setbacks to accomplish an important challenge.	❑	❑	❑	❑	❑

P2- Socio-emotional skills (Locus of control)

For each of the following questions, answer **Yes** or **No**:

	Question	**Yes**	**No**
1.	Do you usually feel that it is almost useless to try at school because most children are smarter than you?	❑	❑
2.	When bad things happen to you, is it usually someone else's fault?	❑	❑

B- Well-being and motivation to attend school

P3- For each of the following statements, mark only one of the following options: 1) almost never, 2) sometimes, 3) often, 4) almost always.

		1	**2**	**3**	**4**
1.	I love learning new things in class.	❑	❑	❑	❑
2.	I feel good at my school.	❑	❑	❑	❑
3.	I feel that the things I do at school are important.	❑	❑	❑	❑
4.	I am a good student.	❑	❑	❑	❑
5.	I am very interested in the things I do at school.	❑	❑	❑	❑
6.	I feel that I can be myself at school.	❑	❑	❑	❑

		1	2	3	4
7.	I think school is important and should be taken seriously.	❑	❑	❑	❑
8.	I do a good job in my subjects.	❑	❑	❑	❑
9.	I enjoy working on class activities and projects.	❑	❑	❑	❑
10.	I do well on my class assignments.	❑	❑	❑	❑
11.	I feel happy working and learning at my school.	❑	❑	❑	❑
12.	I feel treated with respect at my school.	❑	❑	❑	❑
13.	I think what I learn at school will be useful in my life.	❑	❑	❑	❑
14.	I get good grades.	❑	❑	❑	❑

P4- Complete the following statements with one of the following options: 1) very satisfied, 2) satisfied, 3) dissatisfied, 4) very dissatisfied.

		1	2	3	4
1.	My family life.	❑	❑	❑	❑
2.	My friendships.	❑	❑	❑	❑
3.	My experiences at school.	❑	❑	❑	❑
4.	Myself.	❑	❑	❑	❑
5.	Where I live.	❑	❑	❑	❑

C- School stress

P5- For each of the following statements, mark only one of the following options: 1) strongly disagree, 2) disagree, 3) neither agree nor disagree, 4) agree, 5) strongly agree.

		1	2	3	4	5
1.	I have too much homework.	❑	❑	❑	❑	❑
2.	Every day I have to learn too many things.	❑	❑	❑	❑	❑
3.	Schoolwork is very tiring.	❑	❑	❑	❑	❑
4.	I feel pressure to pass exams in order to move on to the next grade.	❑	❑	❑	❑	❑
5.	My parents expect too much from me at school.	❑	❑	❑	❑	❑
6.	I am quite slow at finishing my homework.	❑	❑	❑	❑	❑
7.	Even though I try hard, I do not get the recognition I deserve from my teachers.	❑	❑	❑	❑	❑
8.	I am always satisfied with the feedback my teachers give on my work.	❑	❑	❑	❑	❑
9.	When I need help, I can get it from my teachers.	❑	❑	❑	❑	❑
10.	As soon as I get up, I start thinking about my study problems.	❑	❑	❑	❑	❑
11.	If I put off something I had to do today, I will have trouble sleeping tonight.	❑	❑	❑	❑	❑

D- Aspirations

P6- What are your plans when you finish compulsory education? (Select one option)

- ❑ Vocational education and training
- ❑ Continue studying (Upper secondary/Bachillerato)
- ❑ Look for a job
- ❑ I don't know

Answer **Yes** or **No**:

- Would you like to go to university in the future? Yes ❑ No ❑
- If so, do you think it would be possible? Yes ❑ No ❑

P7- Time spent doing homework. Thinking about the past month, how much time did you spend on homework on average per day? (Select one option)

- ❑ Less than 15 minutes
- ❑ 15–30 minutes
- ❑ 30–60 minutes
- ❑ 1–1.5 hours
- ❑ 1.5–2 hours
- ❑ 2–2.5 hours
- ❑ More than 2.5 hours

E- Interest in mathematics and reading
P8- How much do you like the following subjects? (Select one option)
Spanish/Catalan language:

- ❑ I like it a lot
- ❑ I like it quite a bit
- ❑ I neither like nor dislike it
- ❑ I like it a little
- ❑ I don't like it

Mathematics:

- ❑ I like it a lot
- ❑ I like it quite a bit
- ❑ I neither like nor dislike it
- ❑ I like it a little
- ❑ I don't like it

G- Self-efficacy and anxiety in mathematics
P9- Answer on a scale of: 1) Never, 2) Almost never, 3) Sometimes, 4) Many times, 5) Usually.

		1	**2**	**3**	**4**	**5**
1.	I have been able to understand mathematics.	❑	❑	❑	❑	❑
2.	I have been good at mathematics.	❑	❑	❑	❑	❑
3.	I have enjoyed mathematics.	❑	❑	❑	❑	❑
4.	I am the kind of person who can learn mathematics well.	❑	❑	❑	❑	❑
5.	I have been happy in mathematics classes.	❑	❑	❑	❑	❑
6.	Mathematics teachers have been interested in helping me learn the material.	❑	❑	❑	❑	❑
7.	I have asked questions in mathematics class.	❑	❑	❑	❑	❑
8.	I have asked mathematics teachers for help outside of class.	❑	❑	❑	❑	❑

		1	2	3	4	5
9.	I have set goals for mathematics.	❑	❑	❑	❑	❑
10.	I have worked with other students in mathematics classes.	❑	❑	❑	❑	❑
11.	I have worked hard in mathematics classes.	❑	❑	❑	❑	❑
12.	I regularly do the mathematics homework I am assigned.	❑	❑	❑	❑	❑
13.	Doing mathematics homework makes me feel stressed.	❑	❑	❑	❑	❑
14.	I worry that I might not be able to understand mathematics.	❑	❑	❑	❑	❑
15.	I get nervous when asking questions in class.	❑	❑	❑	❑	❑
16.	I feel tense when preparing for a mathematics exam.	❑	❑	❑	❑	❑
17.	I think I can pass mathematics in every school year.	❑	❑	❑	❑	❑
18.	I think I am the type of person who is good at mathematics.	❑	❑	❑	❑	❑
19.	I worry that I might not do well on mathematics exams.	❑	❑	❑	❑	❑
20.	I worry that I do not have a strong enough mathematics foundation to do well in future mathematics courses.	❑	❑	❑	❑	❑
21.	I think I can get a high grade in mathematics.	❑	❑	❑	❑	❑
22.	I worry that I might not be able to get a good grade in mathematics.	❑	❑	❑	❑	❑
23.	I think I can learn well in mathematics class.	❑	❑	❑	❑	❑
24.	I think I can think like a mathematician.	❑	❑	❑	❑	❑
25.	I think I can complete all mathematics class tasks.	❑	❑	❑	❑	❑
26.	I get nervous when I have to use mathematics outside school.	❑	❑	❑	❑	❑
27.	I think I can understand what is taught in mathematics class.	❑	❑	❑	❑	❑
28.	I think I can do well on a mathematics exam.	❑	❑	❑	❑	❑
29.	I feel overwhelmed when mathematics teachers explain the lesson.	❑	❑	❑	❑	❑
30.	I worry that I will have to use mathematics in my future career.	❑	❑	❑	❑	❑

A.2 Tutor survey

ONLINE SMALL-GROUP TUTORING PROJECT FOR STUDENTS FROM VULNERABLE BACKGROUNDS OF THE GOVERNMENT OF NAVARRE

QUESTIONNAIRE FOR TUTORS

A- Demographic variables

1. Tutor identifier: _____
2. Age: _____
3. Sex: _____
4. Place of birth (municipality, province, country): _____
5. What is your highest level of education completed? (check one option)
   - ❑ Primary education
   - ❑ Secondary education (incomplete)
   - ❑ Secondary education (completed)
   - ❑ Non-university higher education

- ❑ University higher education (BA/BS, Licenciatura/Grado) Degree (e.g., History, Biology, Architecture): _____
- ❑ Master's degree
- ❑ PhD
- ❑ Other. Specify: _____

6. Occupation prior to the start of online tutoring:

- ❑ Unemployed/Inactive
- ❑ Full-time worker
  Occupation:
  - ❑ Primary school teacher
  - ❑ Secondary school teacher
  - ❑ Other (teaching-related)
  - ❑ Other
- ❑ Part-time worker
  Occupation (select all that apply):
  - ❑ Primary school teacher
  - ❑ Secondary school teacher
  - ❑ Other (teaching-related)
  - ❑ Other

B- Previous experience as a teacher

7. Have you ever worked as a teacher?

- ❑ No (if you select this option, go to question 8)
- ❑ Yes

7.1 How long have you worked as a primary education teacher? [months/years]: _____

If your answer is 0 months, go to question 7.2. What subjects have you taught in the past as a primary education teacher? (Select all that apply)

7.2 How long have you worked as a secondary education teacher? [months/years]: _____

If your answer is 0 months, go to question 7.3. What subjects have you taught in the past as a secondary education teacher? (Select all that apply)

7.3 How long have you worked as a higher education teacher? [months/years]: _____

C- Experience as a tutor
Before online tutoring, had you ever worked as a private tutor (online or in person)?

- ❑ No (if you selected this answer, go to question 12)
- ❑ Yes

If you answered **Yes**:

- How long had you worked as a tutor? [months/years]: _____
- How long had you worked as an online tutor? [months/years]: _____

A.3 Additional Tables

Table A.1: Tutor characteristics

	Obs.	Mean	Std. dev.	Min.	Max.
Wave 1					
Female	35	0.82	0.39	0	1
Age	35	33.92	7.38	24	51
Spanish nationality	35	0.97	0.18	0	1
Master's degree	35	0.66	0.47	0	1
Previously worked as a teacher	35	0.91	0.29	0	1
Previously worked as a math teacher	35	0.84	0.36	0	1
Previously worked as a tutor	35	0.50	0.50	0	1
Wave 2					
Female	44	0.69	0.46	0	1
Age	44	32.46	6.28	24	51
Spanish nationality	44	0.97	0.17	0	1
Master's degree	44	0.58	0.49	0	1
Previously worked as a teacher	44	0.94	0.23	0	1
Previously worked as a math teacher	44	0.86	0.35	0	1
Previously worked as a tutor	44	0.70	0.46	0	1

Notes: The table shows summary statistics for tutor characteristics across Wave 1 and Wave 2 of the program.

Table A.2: Missing values

	(1) Control mean	(2) Treatment coef.	(3) (s.e.)
Academic outcomes			
Maths - standardized test	0.105	-0.055***	(0.016)
Maths grade - school transcripts	0.048	-0.015	(0.010)
Grade average- school transcripts	0.048	-0.015	(0.010)
Socio-emotional outcomes and aspirations			
Self-efficacy & maths anxiety index	0.619	-0.125***	(0.027)
Interest in maths	0.544	-0.170***	(0.025)
School stress index	0.563	-0.152***	(0.026)
School well-being and motivation index	0.777	-0.061***	(0.019)
Grit index	0.566	-0.162***	(0.026)
Plan to study high school	0.544	-0.172***	(0.025)
Aspiration to attend university	0.544	-0.169***	(0.025)
Life satisfaction index	0.558	-0.160***	(0.026)
Interest in language (Spanish/Basque)	0.544	-0.169***	(0.025)

Notes: Column 1 reports the mean for the control group. Columns 2 and 3 present the coefficient and standard error from regressions of the variable reported in column 1 on a treatment dummy. *** p<0.01, ** p<0.05, * p<0.1.

Table A.3: Students characteristics at baseline. Conditional on taking the standardized test.

	(1) Obs.	(2) All	(3) Control	(4) Treatment	(5) P-value	(6) Std. diff.
Students characteristics						
Female	1256	0.512	0.506	0.514	0.379	0.016
Age	1256	11.455	11.554	11.419	0.855	-0.092
5 EP	1256	0.431	0.413	0.437		0.048
6 EP	1256	0.326	0.350	0.317		-0.071
1 ESO	1256	0.138	0.129	0.141		0.036
2 ESO	1256	0.106	0.108	0.105		-0.008
Spanish nationality	1256	0.675	0.671	0.677	0.862	0.013
SESN census	954	0.591	0.592	0.591	0.998	-0.002
School in an urban area	1256	0.157	0.159	0.156		-0.007
School in a semin-dense area	1256	0.331	0.347	0.325		-0.046
School in a rural area	1256	0.512	0.494	0.518		0.049
Academic outcomes						
Maths - standardized test	1256	2.958	2.783	3.021	0.069	0.140
Maths grade - school transcripts	1256	2.311	2.398	2.280	0.039	-0.081
Grade average - school transcripts	1256	2.860	2.877	2.854	0.126	-0.020
Socio-emotional outcomes and aspirations						
Self-efficacy and maths anxiety index	1256	1.448	1.297	1.503	0.533	0.099
Interest in maths	931	3.440	3.469	3.430	0.585	-0.030
School stress index	1256	1.206	0.992	1.283	0.006	0.179
School well-being and motivation index	511	0.314	0.298	0.320	0.775	0.024
Grit index	1256	1.233	1.023	1.309	0.012	0.177
Aspiration to study *Bachillerato*	1256	0.220	0.249	0.209	0.943	-0.093
Aspiration to attend university	1256	0.320	0.359	0.306	0.951	-0.113
Life satisfaction index	1256	1.355	1.172	1.422	0.057	0.142
Interest in language (Spanish/Basque)	924	3.749	3.727	3.757	0.878	0.027
Joint test (p-value)					0.682	

Notes: Column 1 reports the number of observations. Columns 2–4 show the mean for the full sample, the control group, and the treatment group, respectively. Column 5 presents p-values from regressions testing whether the treatment–control mean difference is zero, controlling for stratification and wave fixed effects and using robust standard errors. The last row reports the p-value from a joint test of overall significance from the same specification. Column 6 shows the standardized difference in means, calculated as the difference in means divided by the pooled standard deviation. For a variable X, let $\bar{X}_1$ and $\bar{X}_0$ denote the sample means in groups 1 and 0, and let s_1 and s_0 denote the corresponding sample standard deviations. The pooled standard deviation is $s_{\text{pooled}} = \sqrt{\frac{s_0^2+s_1^2}{2}}$. The standardized difference is then $\text{StdDiff} = \frac{\bar{X}_1-\bar{X}_0}{s_{\text{pooled}}}$. *** p<0.01, ** p<0.05, * p<0.1.

Table A.4: Students characteristics at baseline. Conditional on having school records.

	(1) Obs.	(2) All	(3) Control	(4) Treatment	(5) P-value	(6) Std. diff.
Students characteristics						
Female	1302	0.510	0.510	0.510	0.377	0.000
Age	1302	11.482	11.527	11.466	0.839	-0.041
5 EP	1302	0.431	0.434	0.430		-0.008
6 EP	1302	0.320	0.330	0.316		-0.030
1 ESO	1302	0.141	0.132	0.144		0.032
2 ESO	1302	0.109	0.104	0.111		0.021
Spanish nationality	1302	0.675	0.673	0.676	0.512	0.005
SESN census	981	0.599	0.604	0.598	0.935	-0.012
School in an urban area	1302	0.159	0.172	0.154		-0.048
School in a semin-dense area	1302	0.336	0.349	0.331		-0.040
School in a rural area	1302	0.505	0.479	0.515		0.073
Academic outcomes						
Maths - standardized test	1302	2.897	2.694	2.973	0.021	0.159
Maths grade - school transcripts	1302	2.319	2.451	2.269	0.024	-0.125
Grade average - school transcripts	1302	2.852	2.901	2.833	0.119	-0.060
Socio-emotional outcomes and aspirations						
Self-efficacy and maths anxiety index	1302	1.398	1.201	1.472	0.419	0.134
Interest in maths	956	3.432	3.469	3.419	0.751	-0.039
School stress index	1302	1.165	0.902	1.264	0.007	0.226
School well-being and motivation index	543	0.308	0.314	0.306	0.659	-0.008
Grit index	1302	1.184	0.947	1.272	0.021	0.203
Aspiration to study *Bachillerato*	1302	0.225	0.254	0.214	0.720	-0.093
Aspiration to attend university	1302	0.327	0.358	0.316	0.740	-0.089
Life satisfaction index	1302	1.305	1.076	1.390	0.046	0.182
Interest in language (Spanish/Basque)	949	3.737	3.736	3.737	0.886	0.001
Joint test (p-value)					0.523	

Notes: Column 1 reports the number of observations. Columns 2–4 show the mean for the full sample, the control group, and the treatment group, respectively. Column 5 presents p-values from regressions testing whether the treatment–control mean difference is zero, controlling for stratification and wave fixed effects and using robust standard errors. The last row reports the p-value from a joint test of overall significance from the same specification. Column 6 shows the standardized difference in means, calculated as the difference in means divided by the pooled standard deviation. For a variable X, let $\bar{X}_1$ and $\bar{X}_0$ denote the sample means in groups 1 and 0, and let s_1 and s_0 denote the corresponding sample standard deviations. The pooled standard deviation is $s_{\text{pooled}} = \sqrt{\frac{s_0^2+s_1^2}{2}}$. The standardized difference is then $\text{StdDiff} = \frac{\bar{X}_1-\bar{X}_0}{s_{\text{pooled}}}$. *** p<0.01, ** p<0.05, * p<0.1.

Table A.5: Students characteristics at baseline. Conditional on answering endline student survey.

	(1) Obs.	(2) All	(3) Control	(4) Treatment	(5) P-value	(6) Std. diff.
Students characteristics						
Female	771	0.523	0.506	0.527	0.456	0.043
Age	771	11.289	11.329	11.278	0.551	-0.037
5 EP	771	0.466	0.476	0.463		-0.028
6 EP	771	0.392	0.412	0.386		-0.052
1 ESO	771	0.086	0.076	0.088		0.043
2 ESO	771	0.057	0.035	0.063		0.129
Spanish nationality	771	0.702	0.706	0.700	0.542	-0.012
SESN census	636	0.607	0.625	0.602	0.691	-0.048
School in an urban area	771	0.209	0.229	0.203		-0.064
School in a semin-dense area	771	0.235	0.235	0.235		-0.002
School in a rural area	771	0.556	0.535	0.562		0.054
Academic outcomes						
Maths - standardized test	771	3.016	2.797	3.078	0.088	0.164
Maths grade - school transcripts	771	2.403	2.529	2.368	0.127	-0.110
Grade average - school transcripts	771	2.939	2.974	2.929	0.249	-0.039
Socio-emotional outcomes and aspirations						
Self-efficacy and maths anxiety index	771	1.582	1.573	1.585	0.660	0.006
Interest in maths	621	3.576	3.669	3.550	0.442	-0.098
School stress index	771	1.246	1.090	1.291	0.076	0.120
School well-being and motivation index	344	0.340	0.404	0.318	0.345	-0.095
Grit index	771	1.353	1.200	1.397	0.072	0.118
Aspiration to study *Bachillerato*	771	0.228	0.235	0.226	0.090	-0.021
Aspiration to attend university	771	0.357	0.412	0.341	0.903	-0.146
Life satisfaction index	771	1.478	1.417	1.495	0.605	0.043
Interest in language (Spanish/Basque)	619	3.808	3.683	3.844	0.101	0.145
Joint test (p-value)					0.383	

Notes: Column 1 reports the number of observations. Columns 2–4 show the mean for the full sample, the control group, and the treatment group, respectively. Column 5 presents p-values from regressions testing whether the treatment–control mean difference is zero, controlling for stratification and wave fixed effects and using robust standard errors. The last row reports the p-value from a joint test of overall significance from the same specification. Column 6 shows the standardized difference in means, calculated as the difference in means divided by the pooled standard deviation. For a variable X, let $\bar{X}_1$ and $\bar{X}_0$ denote the sample means in groups 1 and 0, and let s_1 and s_0 denote the corresponding sample standard deviations. The pooled standard deviation is $s_{\text{pooled}} = \sqrt{\frac{s_0^2+s_1^2}{2}}$. The standardized difference is then $\text{StdDiff} = \frac{\bar{X}_1-\bar{X}_0}{s_{\text{pooled}}}$. *** p<0.01, ** p<0.05, * p<0.1.

Table A.6: Program participation

	(1) Obs.	(2) All	(3) Wave 1	(4) Wave 2
Absence	971	0.696	0.756	0.658
Number of absences	971	4.135	4.661	3.795
Justified absence	971	0.007	0.016	0.002
Number of justified absences	971	0.012	0.016	0.010
Late arrival	971	0.055	0.060	0.051
Number of late arrivals	971	0.105	0.139	0.083

Notes: Column 1 reports program participation for the pooled sample across both waves. Columns 2 and 3 show the mean participation separately for wave 1 and wave 2, respectively.

Table A.7: Lee bounds

	Maths standardized test			Maths school transcripts			Self-efficacy & anxiety index		
	Lower bound	ITT estimate	Upper bound	Lower bound	ITT estimate	Upper bound	Lower bound	ITT estimate	Upper bound
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Tutoring	0.075	0.200*	0.304***	0.210***	0.210***	0.210***	0.006	0.271***	0.649***
	(0.102)	(0.104)	(0.100)	(0.051)	(0.051)	(0.051)	(0.104)	(0.102)	(0.103)
Observations	1204	1256	1208	1302	1302	1302	512	616	512

Notes: This table presents Lee (2009) bounds for intention-to-treat estimates to address potential sample selection bias arising from differential attrition between treatment and control groups. The bounds are constructed by trimming observations from the group with lower attrition so as to equalize effective sample sizes across groups, yielding upper and lower estimates that bracket the true treatment effect under the assumption of monotonic selection into attrition. All specifications include strata fixed effects. Robust standard errors are reported in parentheses.