Recording the Impact of Social and Emotional Learning*

Salome Aguilar Llanes, Bernardo García Bulle and M. Fernanda Albo Alarcón†

September 28, 2026

[Click here for the latest version]

Abstract

Social skills have been shown to predict earnings, teamwork, and productivity, and they are in increasing demand in the labor market. Schools play a central role in fostering these skills through their regular activities, but also increasingly, through social and emotional learning (SEL) interventions. Despite the widespread adoption of SEL programs, evidence on their effects on classroom dynamics is limited, as most studies rely on self-reports, survey measures, or classroom observation. We implement a randomized controlled trial (RCT) in an online setting where university students tutored elementary and middle school students in math. We randomized the delivery of a structured SEL curriculum of ten one-hour modules, compared to a control group receiving ten one-hour history modules. The SEL group reported improved rapport between tutors and students and increased math learning by 0.0848 standard deviations. To study its impact on classroom dynamics, we collected and transcribed more than 11,700 hours of tutoring sessions and applied a novel AI-based hypothesis generation method to the text. This analysis reveals that SEL-treated math classes featured a slower pace, less technical language, less frustration, more personal topics, and a less serious overall tone in the opening minutes of class. We benchmark these findings against prespecified outcomes from the prior SEL literature — engagement, classroom climate, and instructional quality — and show that the generated hypotheses capture additional variation beyond what these known measures explain.

*We are deeply grateful with Frank Schilbach, Esther Duflo, and Sendhil Mullainathan, for their advice, guidance and support throughout this project. For very helpful discussions and suggestions, we would also like to thank Josh Angrist and James Ross and participants of various internal seminars at MIT. We thank Gabriela Curiel, Maria Maus, Iván Mendoza, and Sebastian Aguilar for their work running this RCT. We also want to thank Cimenna Chao for developing the SEL modules used in this project. AEA registration ID AEARCTR-0009947. MIT COUHES IRB protocol 2101000304 This research project was funded thanks to the J-WEL Grant in pK-12 Education Innovation, the Weiss Fund for Research in Development Economics pilot grant, and the George and Obie Shultz Fund at MIT.

†Aguilar: MIT, saloagui@mit.edu. García: MIT, bernard0@mit.edu, Albo: CMU malboala@andrew.cmu.edu.

1 Introduction

Schools are increasingly recognized not only as places where children acquire academic knowledge, such as mathematics or reading, but also as spaces where they develop the social and emotional skills (SEL skills) that will shape their lives. SEL skills [1] can be defined as "skills that help individuals deal with emotions, including the ability to persist in the face of challenges, manage stress, and get along with others" (Nathanson et al., 2016). A large body of research demonstrates that SEL skills predict important life outcomes, including earnings and educational persistence (Gertler et al., 2014; Walton et al., 2023; Heckman and Kautz, 2012), and enhance desirable outcomes such as teamwork productivity (Weidmann and Deming, 2021). They are also increasingly valued in the labor market: between 1980 and 2012, the share of jobs requiring high levels of social interaction grew by 12 percentage points, while the share of math-intensive jobs declined by 3.3 percentage points (Deming, 2017). The rise of AI is likely to accelerate this trend, making SEL skills an increasingly central component of economic opportunity.

Students can acquire some of these skills through everyday interactions at school. However, the growing recognition of the importance of SEL skills has led most schools to take a more deliberate approach, integrating it into formal instruction and, in some cases, even teaching it as a separate subject. For instance, 83% of U.S. principals report using an SEL curriculum (Skoog et al., 2022), and 90% of schools surveyed by the OECD integrate SEL into teaching (OECD, 2023).

Despite this rise in adoption of SEL interventions, rigorous evidence on whether structured SEL teaching impacts social behavior remains scarce. Most existing studies, while showing experimental evidence, rely on self-reported or teacher-reported surveys (Durlak et al., 2011), which are vulnerable to social desirability bias and awareness of treatment status (Hawthorne effects). Furthermore, measured gains often reflect aggregated survey indices (e.g., self-reported skills, attitudes toward others, social behaviors) rather than observed behavior.

Additionally, while SEL interventions have been widely studied, they differ substantially in content and are often bundled with other programs, such as drug prevention, anti-bullying, or entrepreneurship training, or accompanied by additional resources such as daily visits or supplementary materials. This makes it difficult to isolate the specific effect of SEL content (Heckman and Kautz, 2012; Gertler et al., 2014; Chioda et al., 2021).

Some studies have turned to test scores on subjects such as mathematics or language to examine the impact of SEL interventions (Corcoran et al., 2018; Cipriano et al., 2023; Durlak et al., 2011; Schunk et al., 2022). While these academic outcomes are of independent interest, improving test scores is not necessarily the primary goal of SEL instruction. If SEL works as intended, its effects should also be visible in classroom interactions: settings where students must navigate relationships with teachers and peers, and where they can immediately apply skills such as cooperation, resolving conflicts, and self-regulation.

Consistent with this view, Hagelskamp et al. (2013) study the impact of SEL on classroom climate through direct observations and find improvements across roughly 300 recorded classes. Yet, collecting detailed classroom data is costly and difficult to scale, and until recently this had largely limited the study of classroom dynamics to qualitative research. While such work provides valuable insights, it comes with important limitations: it can only be carried out at small scale, it risks altering the very interactions it

[1] SEL skills are often referred to as personal traits, non-cognitive skills, or soft skills

aims to capture, and it is subject to observer bias. Even when richer data can be collected, it is often difficult to analyze it systematically.

To overcome these constraints, novel technological tools are increasingly being leveraged to facilitate new ways to study the impacts of SEL. Advances in transcription, natural language processing (NLP), and large language models (LLMs) make it possible to analyze large volumes of classroom interaction data, enabling scalable, systematic, and less intrusive study. This shift is beginning to shape empirical work in education, from analyzing student collaboration (Kim et al., 2025), to identifying effective teaching practices (Demszky et al., 2021; Zhang et al., 2025), and developing real-time feedback tools for teachers (Wang et al., 2025; Sawaya et al., 2025; Demszky et al., 2023).

Classroom interactions are inherently high-dimensional and complex, and their multidimensionality could be affected by an intervention in several ways. However, most empirical work to date has focused on a narrow set of prespecified hypotheses guided by theory or expert judgment. While restricting attention to a few outcomes reduces the risks associated with multiple hypothesis testing, it also limits the scope of what we can learn, substantially narrowing the lens through which we study classroom dynamics.

To make progress on these fundamental issues, we study the impact of formal SEL teaching on classroom interactions, combining a novel dataset of thousands of recorded class sessions with new AI-based methods adapted from the hypothesis generation literature (Ludwig and Mullainathan, 2024; Movva et al., 2025). Our approach consists of two main components.

First, we conducted an RCT of a structured SEL curriculum, carefully designed to isolate the effect of SEL content. The experiment took place from September 2022 to July 2023 with 5,011 students ages 8 to 15 in Mexico. This intervention was implemented through our free online tutoring platform, which, as shown in previous work, increased math scores by 0.14 standard deviations compared to a control group that did not receive the intervention (Aguilar Llanes et al., 2025). In this project, tutors were randomized to deliver math combined with either ten SEL modules, developed by Prof. Cimenna Chao as an adaptation from the established RULER curriculum (Hoffmann et al., 2020), or ten history modules, with history serving as the control. This design ensured that both treatment and control tutoring groups spent equal time together and math instruction remained constant. The groups of students were then randomly assigned to the tutors. Each module required the tutor to complete an asynchronous training that explained the topic and provided a guided activity. Students completed a baseline evaluation before being assigned to a tutor and an endline evaluation in the final two weeks of tutoring.

As a second component of our study, we analyze over 11,700 hours of recorded tutoring sessions collected through our online platform. We adapt the hypothesis-generation framework of Ludwig and Mullainathan (2024) to text, combining it with modern machine learning methods to analyze our unique dataset of online classroom recordings. Rather than restricting our analysis to prespecified outcomes, by generating and testing hypotheses directly from the data, we are able to uncover treatment effects on interaction patterns that would otherwise be missed.

We find three main results. First, the intervention increased students' math scores by 0.0848 standard deviations ($p = 0.0357$). Math scores were measured with a platform-based assessment. This effect is notable because the intervention targeted SEL skills rather than math learning goals, suggesting a spillover effect of SEL on learning.

Next, we analyze the transcripts of the classes and demonstrate that the treatment altered the class

dynamics. We begin by benchmarking our analysis against established outcomes known to be impacted by SEL interventions (Cipriano et al., 2019; Hagelskamp et al., 2013): engagement, classroom climate, and instructional quality. Engagement is proxied by features such as the number of tutor–student exchanges, the time between speaker turns, and the number of questions and exclamation marks in the transcript, combined into an index that increases by 0.13 SD ($p < 0.025$). Classroom climate, which captures the prevalent emotions and rapport between tutors and students, is also positively affected.

On the instructional quality instrument, we observe that math classes in the SEL condition have tutors spend less time on formal mathematical content at the beginning of class, and they score lower on the Mathematics Quality of Instruction (MQI) rubric both in the opening minutes of class (minutes 0–12) and later in the session (minutes 12–25). As we discuss in detail in Section 5, this reflects a difference in lesson structure: SEL tutors appear to invest the opening minutes in building rapport, tend to hold longer sessions, and speak for a larger share of class time. The math learning gains documented in Section 4.1 confirm that this restructuring translated into meaningful gains in student learning, and suggest that the MQI rubric does not capture the dimensions of instruction that drive learning in our setting.

Moreover, our hypothesis generation approach captures variation that traditional rubrics miss. We show that treatment assignment can be predicted using the class transcripts alone. To show this we construct a machine learning predictor using text embeddings, a numerical and vectorial representation of text. Being able to predict the assignment to treatment based on the text transcripts of math classes alone indicates that the intervention left systematic changes on how these sessions unfolded.

The strongest signal appears in the first minutes of class, and we therefore focus our analysis on this portion of the session. We find that SEL-treated classes are characterized by a slower pace, more personal topics, less technical content, less frustration, and less serious interactions in the opening minutes. While some of these findings align with traditional metrics, personal conversation with increased rapport, and less frustration with a more positive classroom climate, others, such as pace and language complexity and the pace of class, are not captured by existing rubrics. Together, these data-driven findings offer a deeper understanding of the mechanisms through which SEL interventions reshape the learning environment.

This paper contributes to several strands of literature.

First, we add to the evidence on the impact of SEL interventions on academic outcomes. Existing randomized studies show that SEL interventions can improve test scores in math and other subjects (Corcoran et al., 2018; Durlak et al., 2011). These interventions vary widely in intensity and length, ranging from as little as 6-120 minutes to as many as 6-180 sessions. Some studies report effects on math scores of 0.1-0.23 SD. Long-term positive effects of SEL are usually found when SEL is bundled with early childhood programs or anti-bullying efforts (Heckman and Kautz, 2012; Gertler et al., 2014; Chioda et al., 2021). While most existing studies focus on in-person classroom interventions, our study evaluates an SEL intervention delivered online, demonstrating that the benefits of these interventions extend to online learning environments.

Second, we contribute to research on the impact of SEL on classroom interactions. We collect a unique dataset of more than 11,700 hours of recorded tutoring sessions, spanning both pre- and post-intervention periods. We show that SEL improved observed rapport measures and that these effects are present even in an online setting. Importantly, we use large language models (LLMs) to analyze classroom transcripts, which allows us to examine thousands of classes rather than relying solely on human observers.

Third, our central contribution lies in adapting a hypothesis generation approach, following Ludwig and Mullainathan (2024), to the study of language in classroom transcripts, showing it provides a powerful new lens for generating and testing hypotheses about skill formation and classroom dynamics. In doing so, we contribute to a growing literature on hypothesis generation from text and high-dimensional data (Ludwig and Mullainathan, 2024; Movva et al., 2025; Gentzkow et al., 2019), extending these methods to the study of educational interventions and demonstrating their value in settings where the relevant dimensions of treatment effect are not known in advance.

The remainder of this paper is structured as follows. Section 2 describes the setting, the tutoring program, and the SEL intervention. Section 3 presents the RCT design. Section 4 reports the intervention's impacts on math scores and survey outcomes. Section 5 introduces the hypothesis-generation approach, which allows us to remain agnostic about specific outcomes. Section 6 concludes with limitations and directions for future research.

2 Setting

Our study takes place in an online tutoring program in Mexico, created and operated by the research team since 2021. The program connects university students with elementary and middle-school students for live online, small-group math instruction. This setting is valuable for studying Social and Emotional Learning (SEL): it combines the flexibility of a scalable online platform with the intensity of repeated small-group interactions. In what follows, we describe the structure of the program, the SEL intervention we implemented, and the timeline of data collection for the RCT.

2.1 The online tutoring program

The tutoring program, called "Jóvenes Ayudando a Niñas y Niños A.C.," (JANN) was initially set up in Mexico in 2021, as a RCT in response to the pandemic to test whether online tutoring could support children's learning. Tutoring interventions have, since the pandemic, become a key resource in education policy worldwide. Tutoring and mentoring are highly effective interventions that can have long-term outcomes on students (Carlana and La Ferrara, 2021; Falk et al., 2020; Angrist et al., 2020; Resnjanskij et al., 2024; Nickow et al., 2020). In our program, we measured an impact of 0.14 standard deviations (SD) on math learning (Aguilar Llanes et al., 2025). This setup has since been used to implement other RCTs, and in 2023, the organization was incorporated as an NGO in Mexico.

The tutors in the program are unpaid volunteer university students. In Mexico, university students are required to complete 480 hours of community service to graduate. They can complete this requirement by working at a civic organization or within their university. JANN partners with universities to allow their students to meet this requirement by serving as volunteer tutors.

The recipients of the tutoring are elementary and middle school students in public schools in Mexico. JANN forms partnerships with state governments, including Aguascalientes, Baja California, Durango, Mexico City, Puebla, Sinaloa, the State of Mexico, and Yucatán, to promote the program among school supervisors and teachers through online information sessions. The teachers then invite parents to sign up their children.

The program uses rolling enrollment, allowing new tutors and students to join throughout the academic year. Each week, we form new groups and assign treatments based on tutor and student availability. Tutors can request to work with one to four groups. Students are randomly assigned to groups of one to five students, with group size determined by tutor and student availability at the time of assignment. We then randomly assign these groups to tutors who are waiting for a group. We provide further details about the randomization in the next section.

Online sessions are held by tutors twice a week, each lasting an hour, and are conducted remotely. Groups meet at the time and day of the week that they agree on. They also establish the platform that they are going to use (Google Meets, Zoom, Microsoft Teams, etc.). Tutors may volunteer from a period ranging from 12 to 24 weeks, depending on the agreement with their university. For each session, tutors are required to upload attendance reports, general information about the class they taught, and its recording. All participants provide consent for these recordings to be used for research purposes. As the final two weeks of tutoring approach, endline surveys are automatically assigned to all participants in each group. We follow up with participants via WhatsApp to ensure they complete these surveys.

2.2 The SEL intervention

Social and Emotional Learning (SEL) skills capture the capacities that allow individuals to recognize and manage emotions, build positive relationships, and make responsible decisions. Multiple frameworks exist for conceptualizing these skills. In this paper, we draw on the Collaborative for Academic, Social, and Emotional Learning (CASEL) framework, which organizes SEL into five core domains:

- **Self-Awareness**: Recognizing emotions, setting goals, and cultivating optimism.
- **Self-Management**: Regulating emotions, managing stress, demonstrating perseverance, delaying gratification, and controlling impulses.
- **Social Awareness**: Showing compassion, understanding social norms, and recognizing support systems.
- **Relationship Skills**: Communicating effectively, cooperating with others, resolving conflicts, and engaging in active listening.
- **Responsible Decision-Making**: Evaluating consequences and making constructive choices.

These skills are closely related to the Big Five personality traits framework (Walton et al., 2021) and are often discussed in the literature on non-cognitive or "soft" skills (Heckman and Kautz, 2012). By developing SEL competencies, students gain tools that not only support learning outcomes but also promote positive social interactions and adaptive behaviors both inside and outside the classroom.

The Social and Emotional Learning (SEL) intervention used in this project is an adaptation of the well-known RULER approach, which stands for Recognizing, Understanding, Labeling, Expressing, and Regulating emotions (Hoffmann et al., 2020). This adaptation was specifically developed for our setting by Professor Cimenna Chao, an academic in the fields of Psychology and Education at the Universidad Iberoamericana in Mexico. The intervention comprises ten modules, described in detail in Appendix

A.1, covering topics such as identifying emotions, practicing breathing techniques, and developing self-awareness.

3 Experimental Design

Our RCT is designed to evaluate the causal impact of the SEL intervention described above. We carefully structured our experiment to hold constant both the total time tutors and students spent together and the instructional time devoted to math. We collected multiple measures, including math test scores, surveys of tutor-student rapport using the Student-Tutor Attachment Survey ((Ang et al., 2020)), and recordings of online session interactions, resulting in a large and unique dataset on learning and social dynamics.

3.1 Experimental design and timeline

We designed our experiment with three broad goals in mind. First, we aimed to estimate the effects of a social and emotional learning (SEL) intervention on observed classroom interactions. To precisely measure this, we asked tutors to collect recordings of all or most of their sessions. Second, we were interested in estimating the effect of the SEL intervention on downstream learning outcomes, in our case, math learning. Third, we also collected a range of additional self-reported measures, including tutor-student rapport and attendance. To test this, we randomized the SEL intervention. Figure 1 provides an overview of the experimental design and timeline of the study.

Enrollment took place on a rolling basis, beginning in September of the 2022/2023 school year and continuing through endline data collection in late 2023. A total of 5,011 students and 781 tutors participated in the RCT. Students signed up to participate in the program and completed baseline evaluations. Tutors registered, completed baseline surveys, an asynchronous training, and an interview. Throughout the fall semester of 2022, students were randomized into treatment groups each Monday, conditional on having at least 150 students ready to participate. Students were first randomly assigned into groups of up to five students with a tutor.

Students and tutors met twice a week for one-hour sessions for 15 to 24 weeks. The duration varied depending on the tutor's university, as each university had different requirements for students to complete their community service. During the teaching period, tutors delivered the ten assigned treatment or control modules. Two weeks before the end of tutoring, students received the endline evaluation, with reminders sent in the following weeks to increase completion rates.

3.1.1 Interventions

- **SEL treatment.** In the SEL group, tutors were required to complete an asynchronous training before implementing it in their tutoring sessions. In the training, they were first introduced to the core concepts of the module, gaining a clear understanding of the skills and objectives being targeted. They were then provided with specific activities to carry out with their students, with each module structured to take approximately one hour to complete. This approach ensured that tutors were prepared both conceptually and practically, allowing them to guide their students effectively through each SEL exercise. Each module was designed to lead students through a progression of

skills, starting with identifying and expressing emotions, moving toward recognizing strengths, practicing regulation, cultivating gratitude, reflecting on challenges, fostering autonomy, and ultimately reframing mistakes as opportunities for growth.

- **Control.** Mexican history content was selected as the control subject because it is unlikely to affect tutor-student rapport or enhance math skills, allowing for a clean comparison of the SEL intervention's effects. Modules in the history control group were designed to closely mirror the SEL modules in structure and format. Tutors completed an asynchronous training that introduced the activities they would carry out with their students, providing the same level of guidance and preparation as in the SEL modules. Each activity was planned to take approximately one hour, ensuring that both groups spent an equivalent amount of time engaged in these activities.

To isolate the effects of SEL, we held two aspects constant across the treatment and control groups. First, the amount of instructional time used to teach math was identical across conditions, ensuring that any observed differences in outcomes could be attributed to the SEL intervention rather than differences in math exposure. Second, the total time tutors and students spent together was also constant, since one of our primary outcomes, rapport (the affective relationship between tutor and student). Rapport can be influenced simply by interaction time and it has been shown to contribute to learning gains, therefore controlling for total time together was important. To satisfy both constraints, we implemented an active control in which tutors in the control group delivered history modules.

Note that in this design, both tutors and students were treated simultaneously. Tutors were themselves randomized into either the treatment or control condition and received training on the relevant content in order to deliver it. Thus, the intervention may have affected outcomes through multiple channels: by changing students' own skills, by altering tutors' skills or behavior, or by shifting the quality of the tutor–student relationship.

3.1.2 Pre-analysis plan

This study was preregistered on the AEA Registry (ID: AEARCTR-0009947). The analysis presented here focuses on the comparison between SEL and non-SEL groups as described in the pre-analysis plan.

Our two main preregistered outcomes are academic learning, measured by student scores on an adaptive math evaluation designed by us, and tutor–student attachment, measured using the student version of the Teacher–Student Relationship Inventory (S-TSRI; Ang et al., 2020). We also preregistered the analysis of features extracted from audio recordings of tutoring sessions. As stated in the pre-analysis plan, we aimed to study "the tone in which the tutor speaks (encouraging, scolding, etc.), the frequency of each student's interventions, among others," with the goal of automating classroom observation using a consistent algorithm to analyze what is happening in each session.

We deviate from the pre-analysis plan in two instances. First, our prespecified design included a cross-randomization with tracking to examine whether peer effects in tracked groups would amplify or dampen the effects of SEL. The analysis presented here focuses on non-tracked students only, and we defer the analysis of the tracking interaction to a separate paper. The point estimates for the tracking and SEL interaction arm are positive (0.05 SD), but we cannot reject the null hypothesis at the 5% level. Prespecified estimates are reported in Appendix Table A.IV. Second, we did not prespecify the hypothesis

generation approach discussed in Section 5. This method addresses multiple-hypothesis testing concerns while remaining flexible about learning from the data. We present some version of the preregistered analysis of the recordings in Section 5.3 and the hypothesis generation approach in Section 5.

3.2 Outcomes

We measured a variety of indicators that can be conceptually grouped into four domains: (i) academic achievement (math scores), (ii) rapport between students and tutors, (iii) administrative metrics, and (iv) classroom interactions. Within each domain (excluding the recordings), our primary outcomes are standardized summary indexes that capture the overall effect of the intervention. We consider individual sub-components, such as specific emotions or individual rapport survey items, as exploratory. Therefore, our key inferences rely on a small set of interpretable aggregate measures.

3.2.1 Math scores

We measure students' academic learning using math scores from an adaptive exam designed for this study. Students completed the exam at both baseline and endline. The exam begins with 10 multiple-choice questions. Correct responses trigger more difficult multiple-choice and open-ended questions, so exam length varies between 15 and 30 minutes. Difficulty also scales with grade level: higher grades face more advanced content, but all versions include the same 10 baseline questions. This ensures a reliable differentiation between students at all skill levels.

Exams were activated in students' profiles at signup, as a prerequisite for tutor assignment, and again two weeks before the end of the sessions. Completion of the endline exam was required to receive a participation diploma, and students were reminded via WhatsApp and incentivized through a lottery. In total, 65% of students in our sample completed the endline math evaluation. To account for differences in age-appropriate content, separate exam versions were administered by grade, and scores were standardized within version.

3.2.2 Student-tutor attachment: rapport

We understand rapport as the quality of the interpersonal relationship between tutor and student. Rapport has been increasingly identified by the literature as a key driver of learning (Cornelius-White, 2007), especially in tutoring contexts where interactions are more personalized (Nickow et al., 2020). In these contexts, rapport develops through sustained exchanges that foster familiarity and trust. Moreover, such a relationship creates an environment prone to risk-taking, asking questions, and sustained engagement (Merriam and Baumgartner, 2020). Rapport also acts as a non-monetary incentive, and is associated with lower dropout rates and higher intrinsic motivation (Fryer Jr, 2016). In this sense, the student-tutor attachment can actively shape learning outcomes.

To measure the student-tutor attachment we used the student version of the Teacher-Student Relationship Inventory (S-TSRI) (Ang et al., 2020). The instrument comprises three dimensions, with each dimension rated on a scale from 1 to 5:

(1) **Closeness:** This dimension assesses whether the student enjoyed the class with the tutor and whether they anticipate missing the tutor.

(2) **Conflict:** This dimension explores whether the student felt frustrated with the tutor.

(3) **Instrumental Help:** This dimension focuses on whether the student would trust the tutor with a problem at home and whether they would seek advice from the tutor.

Each dimension is scored by summing the corresponding items; resulting in three separate scores. An overall attachment index was constructed by averaging these three scores.

3.2.3 Administrative reports and attendance

We collected data from reports that tutors completed after each class. These reports served two purposes: to monitor the implementation of the RCT by tracking issues such as scheduling conflicts, student or tutor absence, and delays in delivering the assigned modules; and to provide outcome measures for the study.

Each report recorded general information about the tutoring session, including the math topics covered, the tutor's experience during the lesson, whether a treatment or control module was delivered, and attendance for each student.

Tutors in the program participate on a volunteer basis and are assigned a specific number of classes to teach. Their activities are monitored throughout the semester to ensure compliance. Nonetheless, not all tutors complete all their assigned classes due to various reasons, which can lead to differences in the number of classes delivered across treatment and control groups. This variability is documented and reported as part of our outcomes.

3.2.4 Recordings

One of our primary outcomes consists of recordings of the tutoring sessions analyzed in Section 5. The recordings were collected by the tutors themselves. After each session, tutors uploaded both a report and the audio recording to the NGO's platform. Tutors were instructed to upload recordings of their classes in order for them to be credited toward their community service requirement. They were allowed a 20% tolerance in case they forgot to record the class, or were unable to teach due to extraordinary circumstances. However, in practice, some tutors stopped teaching once they reached that 80% threshold. Additionally, our dataset is missing some class recordings due to operational issues, such as tutors forgetting to record sessions, students not attending, or corrupted files uploaded to the platform. Due to the project's limited funding, tutors used a variety of software and devices to capture the sessions, resulting in heterogeneous data quality.

Of the 37,369 classes that tutors were assigned to teach during this intervention, 30,081 (80.5%) were reported by tutors as having been taught. Of these, 25,960 had at least one student connected, 24,784 had a file attached, and 20,486 were successfully processed and transcribed. This means that for approximately 78.9% of tutor–student meetings (attended classes), we have a corresponding transcript. Of those, 11,731 correspond to math classes only, which constitute the analytic sample used throughout the analysis (see Appendix Table A.I).

We processed the available recordings as follows. Audios were transcribed using WhisperX (Bain et al., 2023), an open-source automatic speech recognition model developed by OpenAI. This produced a dataset organized at the utterance level, with speakers labeled sequentially as Speaker 1, Speaker 2, and so on, a

process commonly referred to as diarization in the computer science literature. While diarization is not perfect and could be further refined, the results below show that even these noisy measures reveal systematic differences in class dynamics between the treatment and control groups. Additionally, the diarization provides timestamps for each utterance, enabling detailed analyses of the classroom interactions.

3.3 Study population and balance checks

3.3.1 Eligibility criteria and selection

Parents registered their children through the NGO's tutoring platform and provided consent to participate in the research project and for their children to receive online math tutoring free of charge. Students had to be in grades 3rd to 6th of primary school or in middle school in order to be eligible to participate. In addition, students and parents were required to complete baseline surveys to be included in the study.

To join the study, prospective tutors first had to complete a group interview with JANN staff and then undergo structured asynchronous training that covered both pedagogical and subject-specific skills. They also needed to consent to participate in the research study.

3.3.2 Informed consent

Parents and tutors completed a detailed informed consent form collected in their profiles on the platform. Students completed an assent form. Only individuals who completed the consent and assent procedures were included in the study. The consent included information about the tutoring sessions, other experimental measures and surveys, the collection and analysis of the recordings, and the randomized treatments. The specific research hypotheses were not shared with participants to avoid demand effects, but participants were told the range of subjects that could be covered. Instead, the goal of the research was described as work to understand how to "assess whether a tutoring policy would be scalable in Mexico and which would be the best way to design it," and "We will have several groups with different activities. In some groups, we will ask the tutors to use part of the time allotted for tutoring to connect with the students through emotional education or history activities that we will provide. We will intentionally vary the criteria we use to create the groups."

3.3.3 Sample characteristics

Students in the program were enrolled in elementary and middle schools across Mexico, with the majority in eighth grade. Table 1 summarizes key characteristics of the students and their families, as well as balance tests across treatment arms. Approximately half of the students were female. Participants were predominantly from the bottom three income deciles in Mexico, but generally had access to basic services necessary for engaging in the tutoring program. Average family income was around 7,600 pesos per month (roughly 375 USD), placing participants in the second and third income deciles in Mexico in 2023[2]. Parents in the control group had on average 11 years of schooling and parents in the treatment group had 11.3 years, a difference of 0.3 years. The national average years of education is 10.7 years for their age group, indicating that most had begun but not completed high school. About 23% of parents

[2]Income data was collected during implementation, so values are missing for some participants

were single, and only 9% of students had previously received private tutoring. Around 37% of students had no family members who had attended university, and just 2.5% spoke English. Most students had access to essential services: 87% had Wi-Fi, and 97% had electricity. Only 45% of families owned a computer, many students accessed online classes via mobile phones.

As shown in Table 2, tutors are substantially wealthier than the students they teach, with an average family income of 50,000 pesos (approximately 2,500 USD), placing them in the ninth and tenth income deciles. Nearly all tutors (99%) report having electricity and internet access at home, and 76% own a car. Tutors are on average 23 years old, and 47% are female. They have an average GPA of 89 out of 100, and 60% attended private high schools. At the time of volunteering, 38% of tutors were employed, and roughly 45% reported prior teaching experience.

3.3.4 Balance checks

As is expected given the large number of comparisons, a few statistically significant differences across treatment groups are present in the sample, as shown in Table 1, with a joint F-test p-value of 0.11. Parents in the treatment group are on average 0.332 years more educated. Students are 1.4% more likely to have a room to study, and their families are 3.5% more likely to own a washing machine and 2.7% more likely to own a car, suggesting slightly higher household income. We include these variables as controls in the main specification (Equation 1), denoted by X_i.

3.4 Attrition

The platform's design enables participants to complete the endline math evaluation in a separate module from the SEL survey. This approach proved beneficial, as participants responded two shorter surveys instead of a single longer one. However, not all participants who completed one survey completed the other. Therefore, we present attrition rates separately for each survey outcome.

As shown in Table 3, the attrition rate for the math exam was balanced at 35%. We compare the baseline math exam scores conditional on completing the endline exam column (4). The point estimate is close to zero. Thus, students in the treatment and control groups who take the endline exam perform similarly on average, and there is little risk of sampling bias.

3.5 Empirical framework

We focus on the results of the SEL treatment. Most of our empirical analyses estimate treatment effects on outcomes using variants of the following equation:

$$y_i = \alpha SEL_g + X_i\beta + \delta_t + \epsilon_{ig}, \tag{1}$$

where y_i denotes the outcome for student i, such as the standardized endline math score, attendance, or tutor-student attachment measures; SEL_g is a dummy variable indicating that group g was assigned to receive social and emotional learning (SEL) classes; and X_i is a vector of baseline demographic controls, including the baseline math score. As discussed in Section 3.3, we control for demographic characteristics that show slight imbalances between treatment and control groups.

Standard errors are clustered at the group level. We include fixed effects δ_t for randomization day t, since the intervention was implemented in waves across multiple dates. We also add fixed effects for the number of classes assigned to each tutor, as tutors from different universities were expected to teach a different number of classes.

3.5.1 Multiple-hypothesis testing

Given the number of outcomes analyzed in this study, we address concerns related to multiple hypothesis testing. We construct four standardized summary indices that aggregate related outcomes into broader domains, following the methodology of Anderson (2008). These indices reduce dimensionality and improve statistical power while limiting false discoveries.

4 Impacts on math learning and rapport

We focus on the results of the SEL treatment. These results provide evidence that SEL have an overall positive effect on outcomes. This section tests standard hypotheses proposed in the SEL literature. We focus on the two preregistered primary outcomes described in the pre-analysis plan: math learning and tutor–student rapport. We estimate these using the regression framework described in Section 3.5.

4.1 Impacts on math performance

Table 4 reports the impacts of the SEL intervention on math performance. Column (1) shows the regression results without controls, while the other two columns show estimates that additionally control for baseline math scores and other controls. We observe a statistically significant positive impact of the SEL intervention in all three specifications, with an effect size of 0.0744 to 0.0936 standard deviations. To put this in context, this represents roughly 53% to 67% of the learning gains measured in (Aguilar Llanes et al., 2025), where two months of participation in the program led to average math gains of 0.14 standard deviations.

While this represents a moderate effect in the context of education (Kraft, 2020), it is smaller than the average effect of 0.23 SD reported in large-scale studies testing SEL interventions Corcoran et al., 2018, although the intensity of SEL interventions is overall heterogeneous across studies.

Still the effect is notable given that improving math performance is not the primary objective of the SEL intervention. These results are consistent with the idea that SEL skills could enhance students' learning and may generate spillover effects to other subjects. They also highlight the importance of analyzing classroom dynamics and the impact on interactions.

As reported in Appendix Table A.III, administrative data show no significant differences in either student attendance or the number of classes taught by tutors across treatment and control groups. On average, tutors delivered 25 classes, and students attended 10 sessions. These findings suggest that the observed improvements in math learning are likely driven by changes in the dynamics of the tutoring sessions themselves, rather than differences in exposure or attendance. This further motivates our focus on analyzing classroom interactions to better understand the mechanisms behind these gains.

4.2 Tutor-student rapport

We do not observe a significant impact of the SEL intervention on the overall tutor-student rapport index (-0.002 SD, p=0.951), as shown in Figure 2. However, we find a notable positive effect on the dimension of instrumental help (0.064 SD, p<0.1), which captures whether students feel they can rely on their tutor for personal guidance or advice outside of class. This result provides suggestive evidence that the SEL intervention strengthened the affective component of the tutor-student relationship. Rapport is a central element of SEL, reflecting trust, empathy, and the ability to form supportive relationships, skills that extend well beyond the classroom.

5 Recording the impacts of SEL on classroom interactions

Having established the results of the SEL intervention on students' math learning and tutor-student rapport, we turn to a unique aspect of our study: the analysis of classroom recordings. To our knowledge, our dataset is the largest collection of classroom interactions ever recorded in the context of an SEL intervention; the recordings are described in detail in Section 3.2.4. These recordings provide a window into how SEL skills are enacted in real time and how they shape class dynamics. Data of this type are becoming increasingly important as a research input in education and other fields: recent work has used classroom recordings and transcripts to study teaching quality (Demszky et al., 2021, 2023), to process and analyze multimodal educational data (Perez et al., 2025; Vitale et al., 2025), and more broadly text data has become an increasingly important input for economics research (Gentzkow et al., 2019).

Nonetheless, an intervention could influence many dimensions of classroom settings. While one could restrict the analysis to prespecified metrics widely studied in education settings, doing so would overlook the richness of this dataset, leading us to miss an opportunity to learn from the data.

Rather than imposing a fixed set of outcomes or making an exhaustive list of outcomes to test, we adapt the machine learning method for hypothesis generation developed by Ludwig and Mullainathan (2024). Our objective is to characterize how math classes in the SEL treatment differ from those in the control condition, and to translate those differences into testable, interpretable hypotheses. To do this in practice, we invert the standard regression approach: instead of regressing a long list of outcomes extracted from the recordings against treatment assignment, we allow the recordings to have a flexible representation through text embeddings, which we use to train a machine learning predictor of treatment assignment. The success of this predictor is indicative of a treatment effect on classroom interactions. We then show that a set of reasonable prespecified outcomes fails to fully account for the variation captured by the predictor. Controlling for these prespecified outcomes reduces but does not eliminate the algorithm's predictive power, suggesting that the treatment affected dimensions of classroom interaction beyond those already hypothesized.

All data analyzed in this section are restricted to math-only classes, that is, classes where tutors did not deliver either SEL or history activities, which would evidently differ mechanically between the treatment and control conditions. By focusing on math-only classes, we isolate changes in the instructional environment that reflect the broader influence of the SEL intervention rather than its direct content delivery.

We conclude the section by estimating the treatment effect of SEL on these hypotheses in a held-out

validation sample that was not used during hypothesis generation, providing an out-of-sample test of their empirical relevance.

5.1 Treatment assignment can be predicted from text data alone

In this section we show that we are able to predict treatment assignment using the transcription of classroom dialogue alone. We do this by transforming the text into embeddings (numerical and vectorial representations of text) and fine-tuning XLM-RoBERTa (Hugging Face, 2022), a multilingual embeddings model accessible through Hugging Face. We restrict the sample to math classes, excluding any class in which treatment activities (SEL or history modules) were taught. We transcribe the audio recordings using WhisperX (Bain et al., 2023) as described in Section 3.2.4. Using these clean transcripts, we build a predictor by fine-tuning XLM-RoBERTa embeddings and combining it with a second predictor built from text characteristics. We show that assignment to treatment can be predicted with an out-of-sample ROC AUC of 61%. The ROC AUC measures how well the algorithm rank-orders treated and control observations. A random guess would have an AUC score of 0.5, and a perfect prediction would have an AUC of 1.0. The success of the predictor indicates the presence of a treatment effect on classroom interactions: if the SEL intervention left no trace in how math classes unfolded, the algorithm would have no signal to detect.

The predictor combines two complementary models. The first model captures the semantic content of classroom speech. It fine-tunes XLM-RoBERTa (Hugging Face, 2022), a multilingual pretrained language model, directly on the transcribed snippets. XLM-RoBERTa is pretrained on 100 languages including Spanish, and fine-tuning adapts its representations to the specific task of predicting treatment assignment in our context.[3]

We apply early stopping based on AUC on the held-out test set; this allows us to avoid overfitting. Each iteration of the model during fine-tuning minimizes the distance between the predicted scores and the true labels in the training set. At some point the model becomes very accurate in the training set but stops being accurate in the held-out set, meaning the model is overfitting to the specific data it was trained on and is no longer valid out of sample. We stop training when we reach the highest AUC on the held-out set.

We split the data into a training set and a held-out test set at the tutor level, so that no tutor appears in both sets. This ensures that the predictor captures genuine differences between treatment and control that generalize to new tutors, rather than idiosyncratic differences between the specific tutors assigned to each condition in our sample. We balance the training and test sets by undersampling the majority class, so that each set contains 50% treated and 50% control recordings. Concretely, if we have more recordings in the treatment than in the control, we sample from the treatment the same number of recordings as in the control. An alternative would be to oversample, meaning we sample from the minority class, allowing for redraws until both classes are of equal size. Having a balanced sample ensures that metrics such as AUC and accuracy are informative and that the predictor is not biased toward the majority class.

The second model captures the structural dynamics of the conversation. For each snippet we build a representation of the sequence of turns, where each turn is characterized by features such as who is

[3]The model is trained with a learning rate of 2×10^{-5}, a batch size of 16, and gradient accumulation over 2 steps, for up to 5 epochs.

speaking, the length of the silence before the turn, whether the turn contains a question, and how many words it contains. These features capture the rhythm and flow of classroom interaction, who speaks, when, and for how long. The structural model is a hierarchical Transformer encoder (Vaswani et al., 2017) that reads the full sequence of turns and learns which patterns of interaction are associated with treatment assignment. The predictions of the two models are combined using an AUC-weighted average, which assigns each model a weight proportional to its individual AUC on the test set. Both models are trained on the training set and evaluated on the held-out test set; all results reported in this section are computed on the test set to ensure out-of-sample validity. Figure 3 presents the ROC curve for the ensemble predictor.

This approach has important limitations that bear on how the results should be interpreted. The predictor abstracts away from acoustic features such as tone of voice that may themselves be affected by the intervention and that could improve predictive accuracy if incorporated. While the ensemble predictor trained on the first 4 minute snippets achieves an AUC of 61%, predictors trained on other class segments reached at most 55%, suggesting that treatment effects on math-specific interactions occurring later in the session are more subtle and did not yield conclusive hypotheses under our procedure.

Because we allow the algorithm maximum flexibility in representing both the structure and the content of classroom interaction, it serves as an approximate upper bound on the treatment-related variation that can be recovered from transcribed classroom language alone, given the constraints described above. The signal the predictor captures need not coincide with the dimensions of classroom interaction that prior literature has measured.

As we show in the next subsection, a standard set of prespecified outcomes fails to absorb the variation the predictor detects, suggesting that the SEL intervention shaped classroom dynamics along dimensions not fully characterized in previous work. It is this residual variation that motivates our hypothesis generation approach.

5.2 What we know about how SEL impacts class dynamics.

We select outcomes from the literature across three groups of metrics, guided by the CLASS rubric (Pianta et al., 2008) and its application to SEL interventions (Hagelskamp et al., 2013), as well as other outcomes studied in the literature (Cipriano et al., 2019), and adapt them to our small-group online tutoring context. We do this to benchmark our predictor and understand whether the variation it captures can be explained by what we already know. In this section we discuss the impact of the intervention on each of these outcomes; in Section 5.3 we benchmark our predictor against them. For each outcome we estimate:

$$X_r = \alpha + \beta \, \mathrm{SEL}_r + \epsilon_r, \tag{2}$$

where X_r is the standardized score for outcome X in snippet or class r and SEL_r is the treatment indicator.

Engagement. We construct an engagement index from transcript features: number of speaker turns (diarized by WhisperX), and counts of question marks and exclamation points as well as the length of the class and the number of diarized speakers. These outcomes are constructed at the class level and they are standardized and averaged into a single index. We also record total class length in minutes as a complementary measure of sustained participationin Figure 4, we show that these measures in math

classes are impacted by the treatment.

Math classes in the SEL treatment show higher engagement than control classes. The engagement index increases by 0.131 SD ($p = 0.024$), with treated classes showing more frequent turn-taking, and more questions and exclamations. SEL classes also last approximately 3.1 minutes longer on average (0.169 SD, relative to a mean of 42.8 minutes).

Classroom emotional climate.

Following Hagelskamp et al. (2013), we study classroom climate by coding 4-minute segments 1 through 3 of each class (minutes 0–12) using the OpenAI GPT-4o API. For each segment, the model identifies the single most prevalent emotion among three categories: frustrated, neutral, or satisfied. This yields a binary indicator for each category, for example, a segment classified as satisfied receives a score of 1 for satisfied and 0 for both frustrated and neutral. We also ask the model to rate tutor–student rapport within each segment. As shown in Figure 5, SEL classes display a more positive emotional climate: more satisfaction, and higher rapport scores, consistent with the survey-based findings in Section 3.

Mathematics quality of instruction (MQI). Following Hagelskamp et al. (2013), we evaluate instructional quality using the Mathematics Quality of Instruction (MQI) rubric developed by Center for Education Policy Research, Harvard University (n.d.). Unlike the general CLASS rubric, the MQI is tailored to math instruction and is therefore better suited to our setting. The MQI is widely used in education research and has been employed in large-scale observational studies such as the MET project (Kane et al., 2013). It captures several dimensions of instructional practice and student engagement, including (i) share of time teaching math-related content; (ii) MQI1 Richness of math: linking representations to concepts, quality of explanations, use of multiple procedures or solution methods; (iii) MQI2 Working with students and math: remediation of errors, incorporating students' contributions; (iv) MQI3 Errors and imprecisions: fluency in mathematical language, clarity of instruction; and (v) MQI4 Student practices: engagement, contributions, and reasoning.

We apply the rubric to 4-minute classroom segments (minutes 0–12), rated using the GPT-4o API.[4]

Surprisingly, as shown in Figure 6, math classes in the SEL treatment score lower on the MQI rubric relative to control classes. Specifically, treated classes show lower scores on three of the four MQI dimensions: MQI1 Richness of math decreases by 0.13 SD ($p < 0.001$), MQI2 Working with students and math decreases by 0.12 SD ($p < 0.001$), and MQI4 Student practices decreases by 0.12 SD ($p < 0.001$). The share of time spent primarily on math-related content in the first 12 minutes of class is also lower in treated classes (-0.079 SD, $p < 0.001$), while MQI3 Low error prevalence is higher (0.126 SD, $p < 0.001$), meaning that when mathematical content is discussed in SEL classes it is delivered with greater precision.

These findings likely point to a meaningful difference in how SEL tutors structure their classes rather than a deterioration in instructional quality. The results from Section 4.1 establish that the SEL interven-

[4]To assess the reliability of the GPT-4o ratings, we validated the rubric using a sample of 150 recordings that human observers also rated. The inter-rater reliability (IRR) kappa between GPT-4o and our human raters ranged from 0.13 to 0.30, while the IRR kappa among human raters ranged from 0.15 to 0.35. Kappa values range from -1 to 1, with higher values indicating stronger agreement. For comparison, professional human observers in the MET study exhibited IRR kappa values between 0.27 and 0.46 (Kane et al., 2013). Although the reliability of LLM ratings is lower than that of professional observers and therefore not precise enough to provide feedback to individual tutors, it remains informative for research purposes. By binarizing the categories into high and low ratings, inter-rater reliability increases substantially to 0.50, indicating that GPT-4o and our human observers effectively differentiate between high and low scores; it is only finer distinctions (e.g., high vs. very high) that are not labeled consistently.

tion produced significant gains in math learning, so the lower MQI scores came at no detriment to student outcomes. This could reflect the fact that the MQI does not capture the specific dimension of instruction that drives math learning in our setting and that is most affected by the SEL intervention. While tutors devote less time to math at the beginning of class, the question is whether this carries through to the rest of the session.

In Appendix A.5, we show the same MQI scores for minutes 12–24. The share of time spent on math-related content no longer differs between treatment and control in this later segment, consistent with a picture in which SEL tutors invest the opening minutes in building rapport and delay the start of formal math instruction, but converge to a similar instructional focus once the session is underway. The remaining MQI dimensions, however, remain lower in the SEL condition during minutes 12–24 as well, suggesting that the rubric is not capturing, or is miscapturing, the dimension of instruction that drives math learning gains in our setting.

Why might MQI scores be lower despite positive learning outcomes? The MQI rubric is evaluated on fixed 4-minute snippets, while engagement metrics such as number of exchanges and class length are measured over the full session. The positive engagement effects, more speaker exchanges and longer classes overall, as shown in Figure 4, suggest that SEL tutors may be speaking for longer at each turn before switching. This would explain why we observe more total exchanges at the session level but lower participation within a fixed-length snippet: if turns are longer on average, fewer switches occur within any given 4-minute window.

In particular, we show that tutors are speaking for longer in the SEL condition, increasing the share of speaking time they hold relative to students. To examine this, we look at the ratio of speaking time between the top speaker and the second speaker in each session. This is an approximation, since our transcripts are not diarized by role, we cannot directly identify which speaker is the tutor and which is the student, but the top speaker by speaking time is likely to be the tutor in most sessions. The fact that the number of exchanges is nonetheless increasing suggests that students are still responding and participating, even if the tutor is taking up a larger share of each turn. Tutors may be spending more time explaining or building rapport, and this might bring as a by product a slower pace in class as identified in Section 5.4.

5.3 The algorithm's prediction is not fully explained by known outcomes

The algorithm's prediction is not fully explained by the known outcomes shown in Section 5.2. To assess how much of the variation captured by the predictor can be explained by known outcomes, we regress actual treatment assignment on the predicted treatment probability while controlling for the known outcomes:

$$T_r = \beta_0 + \beta_1 \hat{T}_r + X_r \alpha + \epsilon_r, \tag{3}$$

where $\hat{T}_r$ denotes the predictor score for snippet r and X_r represents the three sets of outcomes presented in this section, taken and adapted to our setting from the prior SEL literature: engagement, classroom climate, and instructional quality. In Section 5.4, we return to this equation adding L_r, the hypotheses extracted from our hypothesis generation procedure, to assess how much additional variation they account for beyond the known outcomes.

With these outcomes and their effects on classroom interactions described, Table 5 presents the benchmarking exercise, where we compare these outcomes' signal to that of the predictor. Column (1) shows that regressing actual treatment assignment against the predicted score from the text-based predictor alone yields an adjusted R^2 of 0.035. Columns (2) through (4) show the adjusted R^2 obtained from replacing the predictor with three sets of prespecified outcomes drawn from the prior SEL and education literatures.

Column (2) includes engagement metrics: speaker exchanges, response latency, question and exclamation counts, and total class length. Unlike the predictor and the other outcomes, engagement metrics are measured at the class level rather than the snippet level. Therefore, the regressions in this table aggregate snippet-level predicted scores to the class level by averaging. The adjusted R^2 from this specification is 0.026, comparable in magnitude to that of the predictor. Column (3) adds a second set of prespecified outcomes: classroom climate variables and the prevalence of emotions during the session, yielding an adjusted R^2 of 0.003, substantially lower than Column (1). Column (4) shows the same specification for the Mathematical Quality of Instruction (MQI) index; and the adjusted R^2 of 0.032 is comparable to that of the predictor, reflecting that instructional quality as measured by the MQI rubric is closely tied to the treatment contrast. We show the same specifications estimated at the snippet level in Appendix A.6; the adjusted R^2 for the predictor is slightly lower at 0.027, as expected given that class-level averaging reduces idiosyncratic noise.

A natural concern is whether the predictor is simply recovering variation already captured by these known metrics. Columns (5) and (6) address this. Column (5) includes all three sets of prespecified outcomes jointly, yielding an adjusted R^2 of 0.069. Column (6) adds the predictor's score to this saturated specification; the adjusted R^2 rises from 0.069 to 0.075. The predictor remains statistically significant even after conditioning on the full set of known outcomes, and the increment in adjusted R^2 it contributes indicates that it has detected signal that the prespecified metrics do not absorb. The known outcomes and the algorithm are capturing meaningfully different sources of variation in classroom interactions.

A complementary way to see the same result is to ask how much of the algorithm's predicted score can be explained by the known metrics. When we regress the predicted score against the full set of prespecified outcomes in Table 6 Column 5, the adjusted R^2 is 0.548, meaning these established measures account for just over half of the variation that the predictor captures.

While the SEL intervention thus affected several established dimensions of classroom interaction, these measures do not fully account for the variation the algorithm detects. The predictor captures something beyond engagement, emotional climate, and instructional quality as currently measured. It is this residual variation that motivates our hypothesis generation approach: there is something in the text that the algorithm has found, and our goal is to make it interpretable.

5.4 Extracting interpretable hypothesis

Having established that the algorithm detects systematic differences between treatment and control classes that are not absorbed by known outcomes, we face a second challenge: understanding what the predictor is actually seeing. Knowing that a black-box algorithm can predict treatment assignment is not informative by itself. We want to extract hypotheses that are interpretable, testable, and capable of informing our understanding of how SEL shapes math instruction, and that could ultimately guide future iterations of

the program.

One way to approach the problem would be to compare transcripts with high predicted SEL scores to those with low predicted scores and try to disentangle the differences. However, direct transcript-by-transcript comparison is infeasible for this purpose. Classes vary by tutor, topic, date, and many other contextual factors, so naïve contrasts would largely reflect irrelevant sources of variation rather than systematic differences attributable to the treatment. We therefore adopt a structured approach, adapting the machine learning hypothesis generation procedure of Ludwig and Mullainathan (2024) to our text-based setting. The procedure works in four steps: we first use the predictor to isolate the transcripts most characteristic of each condition, then construct controlled variations of those transcripts to make stylistic contrasts salient, use a large language model to name the recurring differences, and finally validate the resulting hypotheses on held-out data.

A related approach to hypothesis-generation from text data uses sparse autoencoders (Movva et al., 2025). Sparse autoencoders decompose text representations into a large number of interpretable features by learning a sparse, overcomplete basis for the embedding space. Each feature tends to activate for semantically coherent concepts, and hypotheses are generated by identifying which features differ most systematically between groups. The advantage of this approach is that it operates directly on the latent structure of the embeddings without requiring the generation of controlled variations. Our approach differs in that it constructs explicit counterfactual variations of the text, which allows a language model to name differences in natural language through direct comparison, rather than interpreting abstract latent features. The two methods are complementary: sparse autoencoders are more automated and scalable, while our morphing procedure produces hypotheses that are more directly grounded in the text.

Step 1: Predictor. We build on the predictor described in the previous section. To fix notation, let each snippet r be represented by a numerical embedding and the set of relevant features of the text dynamics $x_r \in R^d$. The predictor is defined as

$$m(x_r) = P\big(t_r = 1 \,\big|\, x_r\big), \tag{4}$$

where t_r indicates whether snippet r comes from a math class in the SEL treatment ($t_r = 1$) or from a control class ($t_r = 0$). The predictor $m(\cdot)$ assigns to each snippet a score reflecting how characteristic it is of the treatment condition.

Step 2: Selecting extreme snippets. From the training set we select snippets with the highest and lowest predicted treatment probabilities. Focusing on these extremes increases the signal-to-noise ratio: high-probability snippets are more likely to contain features characteristic of the treatment, while low-probability snippets are more likely to reflect the control condition. Excluding intermediate snippets reduces the influence of idiosyncratic or spurious variation (such as regional speech patterns, day-of-week effects or unrelated exchanges) that is orthogonal to the treatment contrast. Because the purpose of this stage is hypothesis generation rather than causal identification, balance across groups is less critical here. All hypotheses are validated on held-out data in Step 4.

Step 3: Constructing variations to isolate contrast .

Even among extreme snippets, remaining differences orthogonal to the treatment effect will still be present, since snippets are drawn from different groups covering different mathematical topics, with

different tutors, specific language patterns, and incidental references. To isolate the dimensions that the predictor is responding to and that can be attributed to the treatment effect, we generate multiple *variations* (morphs) for each selected snippet. The aim is to create different versions of the same text that are as similar as possible to each other but cause the predictor to change its predicted score. This makes the comparison easier to interpret: instead of comparing two snippets that differ along many dimensions both related and orthogonal to the treatment, we want pairs that are as close as possible except along the dimensions that drive the predicted score, highlighting those differences and minimizing the rest.

To create the variations we prompt an LLM and score each version using the predictor $m(\cdot)$, which increases the number of comparisons we can make. For each original snippet we retain the highest- and lowest-scoring variations and use an LLM to compare them and describe the most notable differences. This procedure is implemented using the Claude 3.5 API.[5] We present an illustrative example in Table 7.

We repeat this across several snippet pairs and identify the differences mentioned most frequently by the LLM across comparisons. These recurring differences are the features the predictor is most consistently detecting. We then select the most mentioned characteristics and label our data with them for validation. In our example, in Table 7, the variation with the highest predicted treatment probability tends to contain more questions, shorter sentences, and a less technical language than the variation with the lowest probability.

Step 4: Validating the generated hypotheses on held-out data. To verify that the identified differences reflect genuine patterns rather than artifacts of the generation procedure, we label all snippets in the training set according to the presence of each candidate characteristic. Each snippet receives a score from 1 to 6, as illustrated in Table 8. We then estimate, for each hypothesis,

$$L_r = \alpha + \beta t_r + \epsilon_r, \tag{5}$$

where L_r is the standardized label score for snippet r and t_r is the treatment indicator. Hypotheses for which β is statistically significant in the training set are retained and used to label the held-out test set. We then repeat the analysis on the test set, which played no role in the hypothesis generation procedure. This out-of-sample validation ensures that the observed differences generalize to new data and are not driven by overfitting.

Step 5: Quantifying explanatory power. Finally, we assess how much of the variation captured by the predictor is accounted for by the extracted hypotheses. We do this in two ways.

First, we regress the treatment probability assigned by our predictor against both the known outcomes and the new labeled characteristics, using test-set data:

$$\hat{T}_r = \beta_0 + X_r\alpha + L_r\gamma + \epsilon_r, \tag{6}$$

where $\hat{T}_r$ is the predictor score, X_r collects known features from the literature (engagement, class climate, and MQI), and L_r contains the labeled hypotheses. We focus on the adjusted R^2, benchmarking the specification that includes only X_r against the one that also includes L_r. The increment indicates how much of the algorithm's signal the extracted hypotheses succeed in making interpretable.

[5]The choice of model was arbitrary; the procedure could be implemented with any comparable LLM.

Second, we regress actual treatment assignment on the predicted treatment probability, controlling for both X_r and L_r, using test-set data, complementing the table discussed in Section 5.3 with our new hypotheses:

$$T_r = \beta_0 + \beta_1 \hat{T}_r + X_r \alpha + L_r \gamma + \epsilon_r, \tag{7}$$

where $\hat{T}_r$, X_r, and L_r are defined as above. We again focus on the adjusted R^2, comparing the specification that includes only the known features X_r to the one that also includes the extracted labels L_r. A rise in adjusted R^2 indicates that the generated hypotheses succeed in explaining additional variation between treatment and control beyond what existing measures capture.

5.5 Findings: Applying the method to our data

We apply Steps 2 through 5 of the procedure described in Section 5.4 to our classroom recordings. Starting from the text-based predictor, we select the snippets with the highest and lowest predicted SEL scores and generate variations of each. Specifically, we prompt GPT-4o to produce 28 variations of each selected snippet. Initial attempts without detailed instructions produced nearly identical outputs, so we refined the generation process by providing structured guidance: some instructions allowed more flexibility (for example, beginning the conversation similarly but ending differently), while others imposed tighter stylistic constraints. We used GPT-4o to generate the templates for each of the 28 variations and curated the final set of instructions ourselves. We show the prompt in Appendix A.7. This imposed heterogeneity creates a controlled range of stylistic variation while limiting the scope of detectable differences; the results therefore provide a lower-bound estimate of what could be achieved with fully unconstrained morphs.

Following the procedure in Section 5.4, we score all variations using the predictor and construct 200 pairwise comparisons between high- and low-scoring morphs. We feed each pair to the Claude 3.5 API, asking it to name the most salient differences. As with the generation step, prompt refinement was necessary: an initial instruction to identify "differences" led the model to omit some systematic contrasts. We therefore defined explicit categories and instructed the model to identify the three largest differences along each of the following dimensions: interactions, student participation, discussion content, discussion mood, and other differences. The exact prompt is reported in Table A.VII. The resulting word cloud is shown in Figure 7.

From these outputs we extract the eight most frequently mentioned categories as candidate hypotheses. Each description is derived from the qualitative contrasts identified by Claude 3.5 across the 200 comparisons. The eight candidates span multiple dimensions of interactions:

- *Frustration* captures the degree of tutor patience, ranging from very patient to visibly frustrated.
- *Technical content* reflects the technical depth of the dialogue, from simple and accessible language to highly formal and mathematical.
- *Fast pace* measures the rhythm of the session, from too slow to too fast with pressure to keep up.
- *Participation* captures the level of shared turns, from tutor monologue to highly active two-way conversation.

- *Interactivity* reflects the use of questions to promote dialogue, from no questions at all to constant prompting and participation.
- *Confidence* measures the degree to which students display or are encouraged toward confidence, from very timid to actively confident.
- *Personal* captures the warmth and closeness of the relationship, from formal and distant to friendly and rich with personal references.
- *Serious* reflects the overall tone of the session, from relaxed and informal to very serious.

We then apply Step 4 and rate all snippets in the training set on a 1–6 scale using Claude 3.5 guided by each of the eight hypotheses. After standardizing the ratings, we estimate Equation 5 for each label. As shown in Figure 8a, all eight extracted characteristics differ consistently between treatment and control in the training set. Therefore, we carry all eight forward to the test set.

Applying the same labeling and estimation procedure to the held-out test set, we find that five of the eight hypotheses are statistically significant at the 95% confidence level and six at the 90% confidence level. Math classes in the SEL treatment are characterized by less *frustration* (-0.106 SD, $p < 0.001$), simpler and less *technical* language (-0.22 SD, $p = 0.00$), a *slower-paced* , more *personal* topics and references (0.24 SD, $p < 0.001$), a less *serious* overall interaction style and an overall less interactive class. The *confidence* and *participation* labels are not statistically significant in the test set, though their coefficients point in the same direction as in the training data. Results are presented in Figure 8b. This figure compares recordings in the treated and control groups in the test set, confirming that the hypotheses generated by the procedure generalize to the held-out sample. Given that our randomization was conducted at the group level, we also estimate the same specification for the full sample with cluster standard errors to reflect the RCT design. These results are shown in Figure 10 for the full sample; Appendix Figure A.III shows the equivalent results for the test set only. Math classes in the SEL treatment are characterized by less *frustration* (-0.079 SD, $p = 0.060$), simpler and less *technical* language (-0.174 SD, $p = 0.030$), a *slower pace* (-0.190 SD, $p < 0.001$), and more *personal* topics and references (0.265 SD, $p < 0.001$).

Several of these hypotheses map onto the prespecified outcomes examined in Section 5.3, while others point to dimensions of classroom interaction that existing measures do not capture.

The reduction in *frustration* and the decrease in *seriousness* are consistent with the classroom climate and emotion measures reported in Section 5.3. Both also connect directly to the emotion regulation and perseverance skills targeted by the SEL curriculum. Activities 3, 5, and 8 ("Pause to Breathe," "What Can I Do to Regain Calm?," and "Learning from Mistakes") explicitly train students to manage difficult emotions, respond constructively to challenges, and reflect on what can be learned from setbacks. If these skills generalize beyond the SEL session, we would expect tutors and students to respond more patiently and with a lighter tone when mathematical difficulties arise.

The increase in *personal* topics is consistent with the increased rapport documented through the student–tutor attachment questionnaire in Section 4.2. Activities built around gratitude, self-esteem, and personal reflection (Activities 4, 6, and 7) encourage students to share aspects of their lives and feelings with their tutor, naturally bringing a warmer and more personal tone into subsequent math interactions.

The decrease in *technical content* and *fast pace* are not directly captured by existing rubrics or engagement indices, and represent dimensions of classroom interaction that the hypothesis generation procedure reveal that would not otherwise have been tested. These features may reflect tutors developing a closer understanding of their students through the SEL activities and adjusting the language and rhythm of instruction accordingly. Activities centered on self-awareness and sharing (Activities 1, 2, 6, and 7) create opportunities for tutors to learn how students process information and where they struggle, which could translate into simpler explanations and a more measured pace in math instruction.

This contrast between findings demonstrates the complementarity of prespecified measures and hypothesis generation as tools for understanding how SEL interventions reshape the learning environment.

One finding should be further discussed. The *participation* label shows a decrease in shared turns in SEL classes relative to control, which appears to contradict the increase in speaker exchanges documented in Figure 4. These two measures are not directly comparable, however, and the evidence from Section 5.3 helps reconcile them. The speaker exchange count is computed over the full class session, whereas the participation label is derived from fixed-length snippets without speaker diarization. As discussed in Section 5.3, SEL tutors appear to speak for a larger share of class time, consistent with longer individual turns. If turns are longer on average, fewer switches occur within any fixed-length snippet, which would produce exactly the pattern we observe: more total exchanges at the session level but lower participation within a snippet. Importantly, the fact that total exchanges are increasing suggests that students are still responding and participating but the tutor is simply holding the floor for longer stretches before each exchange. This interpretation is also consistent with the lower scoring in the MQI rubric.

Finally, we apply Step 5 on the test set to assess how much of the predictor's signal is accounted for by the extracted labels, estimating Equation 6. Results are reported in Table 6. Regressing the probability score assigned by the predictor against the extracted labels alone yields an adjusted R^2 of 0.51 shown in Column (4), comparable to the adjusted R^2 of 0.548 obtained from the full set of known outcomes, Column (5). When we include both the extracted labels and the known outcomes jointly, Column (6), the adjusted R^2 rises to 0.676, indicating that the generated labels explain a share of the predictor's signal that is not captured by existing measures alone. This shows both that the predictor is capturing a different signal from existing measures and that we are succeeding, at least in part, in making that signal interpretable.

Further evidence comes from estimating Equation 7, regressing the actual treatment assignment on the extracted labels as the only regressors, which yields an adjusted R^2 of 0.043, shown in Column (2), which is comparable to the adjusted R^2 0.035 obtained from the predictor score alone, shown in Column(1) of Table 9. As a benchmark, the known features studied in the literature yields an adjusted R^2 of only 0.069 (Column 5 of Table 5), confirming that the procedure recovers a meaningful signal well beyond what existing measures capture.

These results provide proof-of-concept evidence that the hypothesis generation procedure identifies interpretable and generalizable differences between SEL and control classrooms. However, the procedure is not exhaustive, and several avenues for improvement remain. The morph generation step could be strengthened by allowing more flexible and heterogeneous stylistic variations, and the predictor could be enriched with audio features that carry information about tone and affect. Each of these extensions would likely sharpen both predictive accuracy and the hypotheses the procedure surfaces. Nonetheless, we learnt

that relative to control classes, SEL math classes in the first few minutes are slower paced and feature a lighter load of technical content, suggesting that the intervention reshapes not just the emotional tone of instruction but also how mathematical material is introduced and delivered.

6 Conclusion

Social skills predict important life outcomes, including earnings and educational persistence (Gertler et al., 2014; Walton et al., 2023; Heckman and Kautz, 2012). They are increasingly valued in the labor market (Deming, 2017), a trend likely to be reinforced by the rise of AI. Yet most empirical work relies on self-reported or teacher-reported surveys to study skill formation, leaving limited evidence on actual behavioral changes.

SEL skills are fundamentally social: they shape how individuals regulate emotions, manage challenges, and navigate relationships. However, most evaluations of SEL interventions focus on individual outcomes rather than the social interactions where these skills matter most. Interactions are inherently relational: they emerge from the equilibrium of multiple actors' skills. In schools, more engaged students can motivate teachers to invest greater effort in lesson preparation, while well-prepared teachers can in turn foster higher student engagement. These feedback loops are difficult to disentangle, but they are precisely where learning takes place. In this paper we focused on studying interactions directly as outcomes of an SEL intervention.

We launched a randomized controlled trial in an online tutoring program in Mexico to study how an SEL intervention influenced classroom interactions and student learning. We implemented multiple strategies to analyze treatment effects on interactions and adapted a method to generate hypotheses directly from text data. Our findings show that teaching SEL improved tutor–student relationships and, importantly, increased math learning, even though math was not the primary focus of the intervention.

We collected an extensive dataset of over 11,700 hours of recorded sessions and tested results previously documented in the literature. Math classes in the SEL condition exhibit higher levels of engagement, and emotions during math instruction tend to be more positive, replicating patterns found in prior SEL evaluations. When rating sessions with the MQI rubric, widely used in the education literature, we find that tutors in the SEL condition devote less of the opening minutes of class to formal mathematical content. As we argue in Section 5, this probably reflects a difference in lesson structure rather than instructional quality: SEL tutors invest the early portion of class in building rapport, and the math learning gains establish that this came at no detriment to student outcomes.

We then leveraged a hypothesis generation approach to exploit the richness of the recordings. This method surfaces insights that complement existing measures, identifying that SEL math classes begin at a slower pace, contain more personal interactions, and feature less technical language in the opening minutes, outcomes that go beyond what would have been tested using prespecified rubrics alone. Such an approach is particularly valuable in high-dimensional settings, where the pattern-recognition capabilities of automated predictors can identify and formalize potential treatment effects in a structured, data-driven way.

This study has several limitations. First, it is set in an online tutoring context, which limits our ability to generalize the results to in-person classrooms. Second, our hypothesis generation method relies

on a morphing procedure that produces stylistic variations while holding base content constant. While useful, this approach cannot capture differences such as whether SEL tutors progress more quickly through material or systematically emphasize different topics. Third, producing high-quality variations with LLMs remains challenging: despite efforts to design prompts that generate flexible outputs, variation was limited and further experimentation is needed. Finally, the hypothesis generation analysis is restricted to the first minutes of class, and further exploration is needed to rule out effects on later portions of the session.

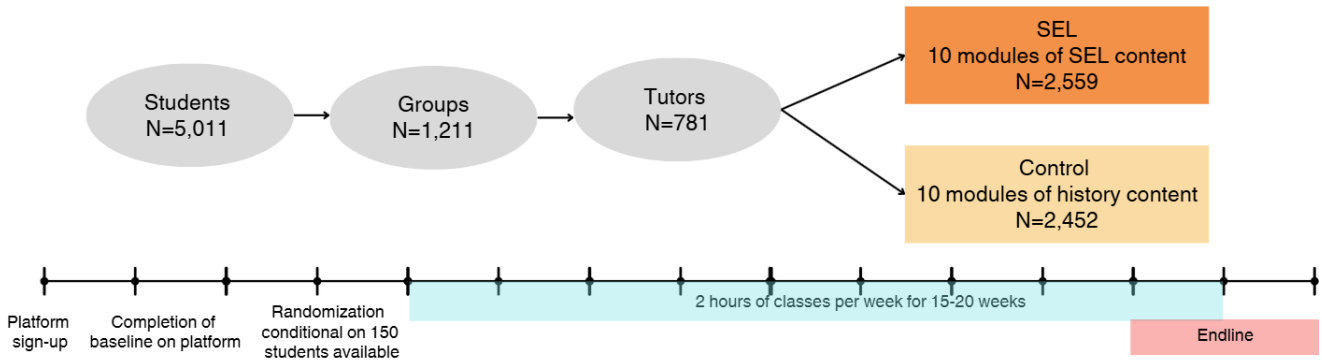
In terms of avenues for future research, the evidence raises the question of what drives the observed impacts. Does SEL primarily enhance (1) students' own SEL skills, (2) the SEL skills of tutors, or (3) the rapport between tutors and students in ways that foster learning? Answering this has direct policy relevance: should SEL be taught by specialized instructors, embedded across all subjects, or delivered through professional development for teachers without reducing curricular time? These questions form the core of a follow-up RCT we are currently implementing, designed to disentangle these channels while leveraging classroom recordings and the hypothesis generation approach introduced here.

The results in this paper illustrate both the challenges and the opportunities of working with large-scale recordings of classroom interactions. We emphasize the richness of these data and provide a proof of concept that they can be used to generate new hypotheses about learning processes. At the same time, substantial scope remains for refining and extending the hypothesis generation approach to better capture the full complexity of classroom text and interaction.

Beyond this setting, hypothesis generation can address other important questions in education research. For example, it could reveal how high-value-added tutors differ in their teaching relative to low-value-added tutors, and what practices explain these differences, or how tutors with traditional gender beliefs interact differently with students of different genders and how shape girls' learning outcomes. Such questions highlight the versatility of the method for uncovering systematic differences in teaching styles and classroom dynamics and their consequences on learning.

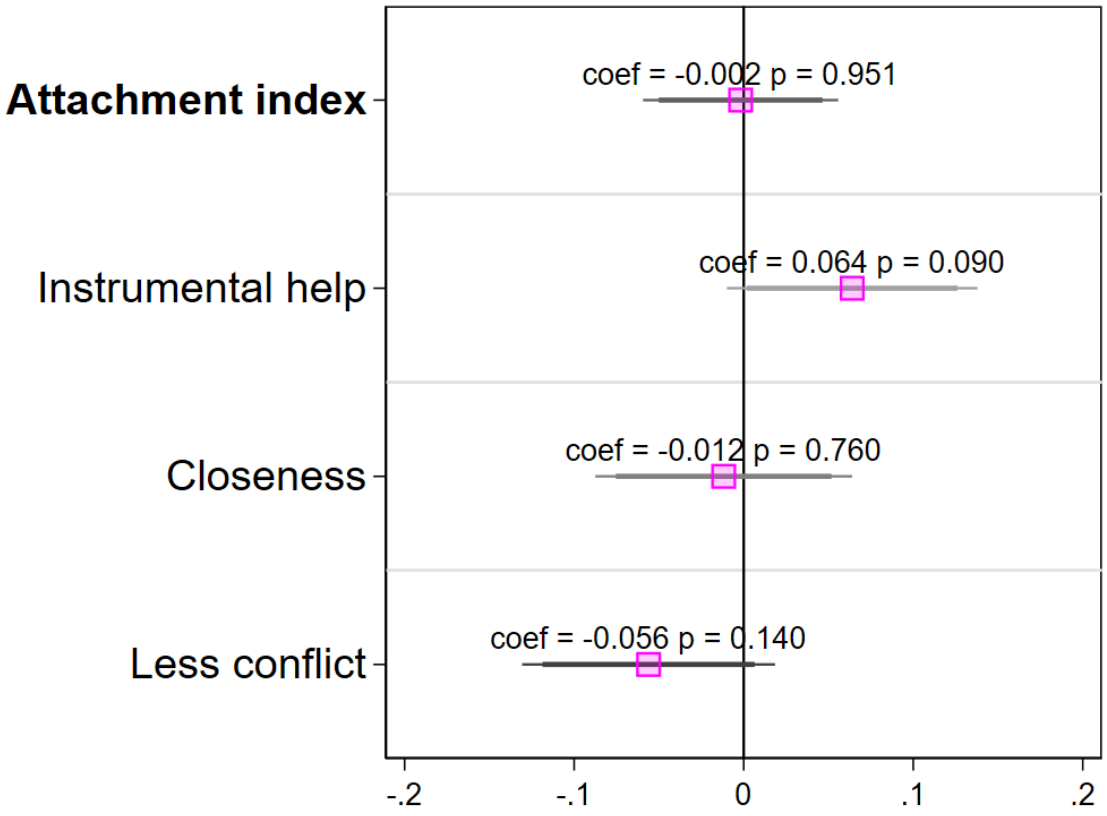
Figures

Figure 1: Experimental Design



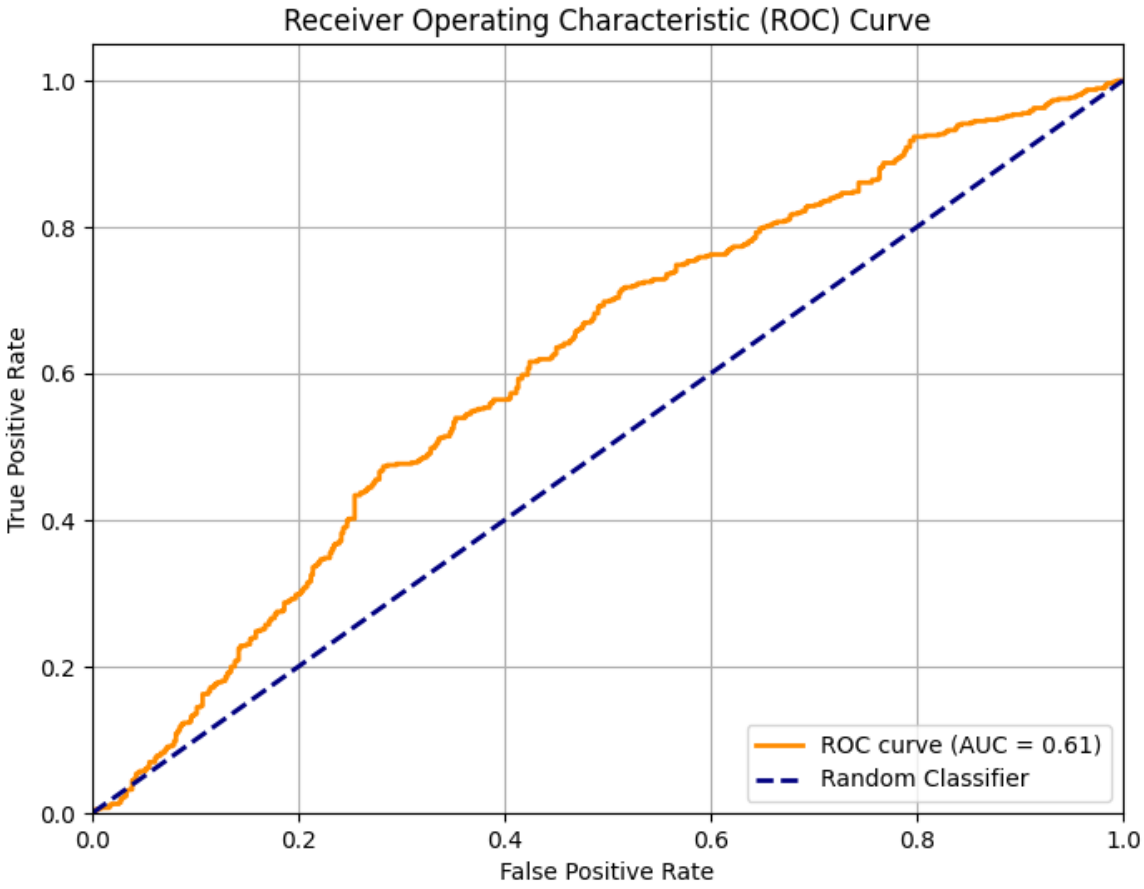
Notes: This figure illustrates the RCT design used in our study.
Students enrolled on the NGO website and completed a series of initial surveys. Enrollment was conducted on a rolling basis, and once enough students were available, they were randomly assigned to groups of up to five students and allocated to tutors. Tutors were tasked with teaching math classes, devoting 10 out of the total classes to either SEL (treatment) or history (control) instruction. The primary comparison in our analysis focuses on outcomes between the SEL and history groups.
Tutors signed up, completed baseline surveys and initial training, and were interviewed before being assigned groups to teach.

Figure 2: Effect of SEL Treatment on Student-Teacher Attachment



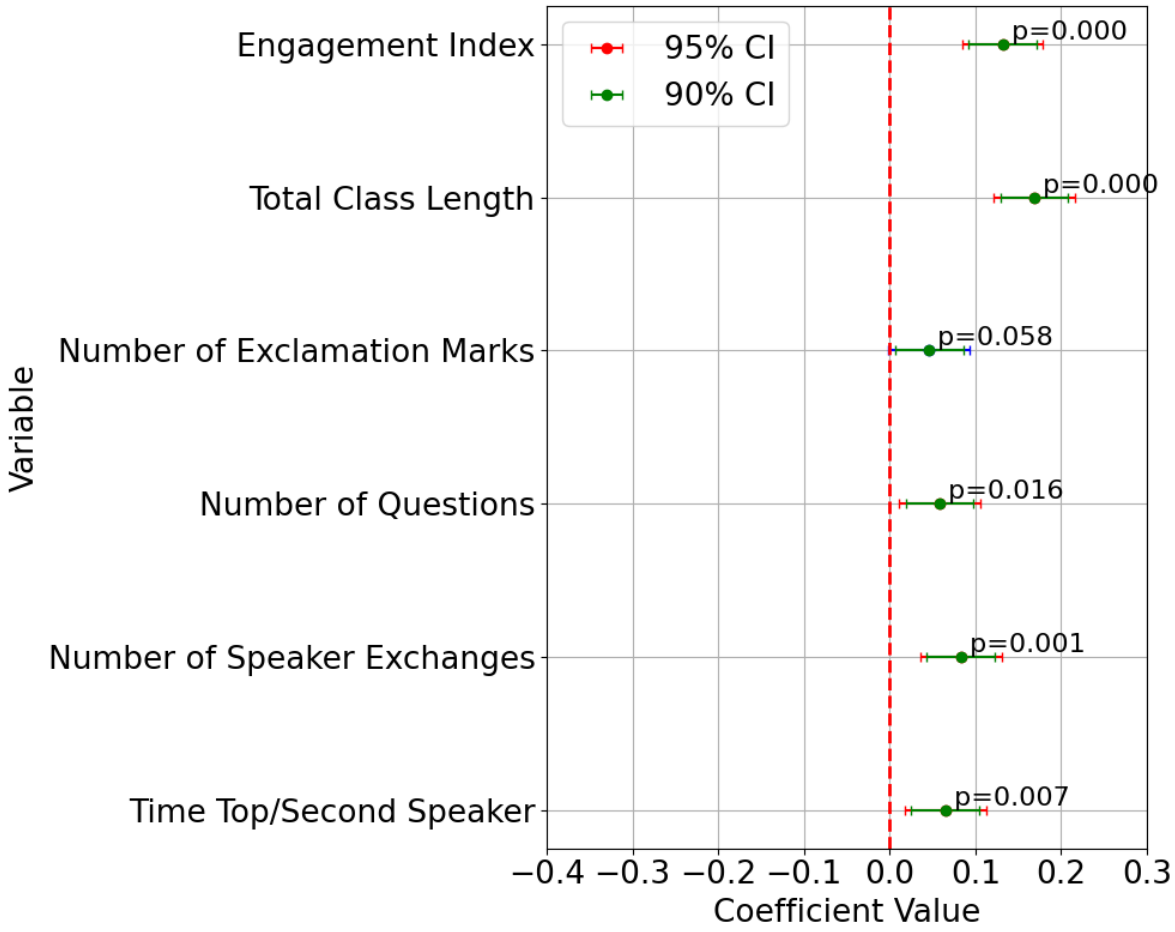
Notes: This figure shows how the SEL treatment (vs. control history lessons) affects dimensions of the student-teacher attachment questionnaire estimating Equation 1. Bars indicate 95% and 90% confidence intervals.

Figure 3: Predictor's ROC Curve



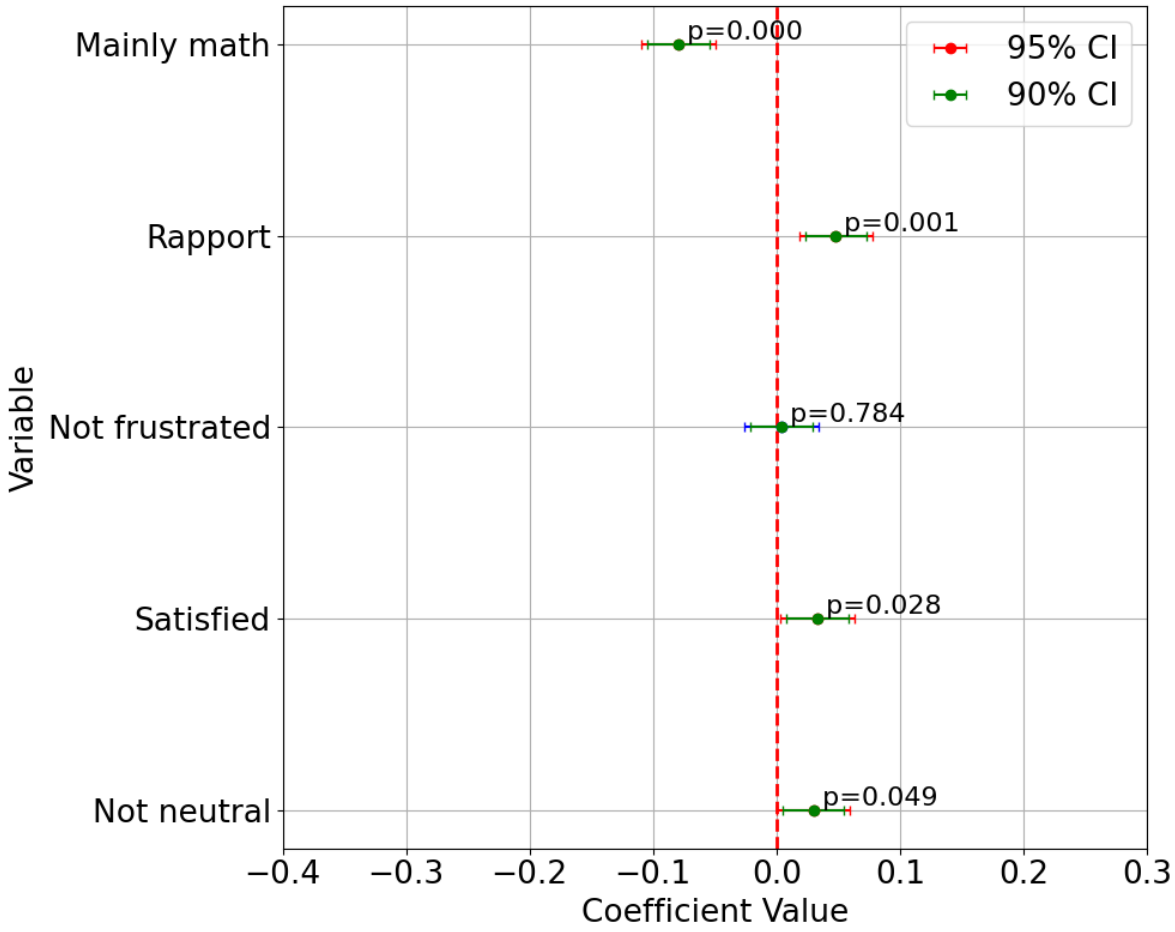
Notes: This figure shows the ROC curve for our predictor. The ROC (Receiving Operating Characteristic) curve is a graphical representation of the performance of a binary classification model across various thresholds. It plots the True Positive Rate (TPR), also known as sensitivity or recall, against the False Positive Rate (FPR), which is (1 – specificity). Specificity stands for the True Negatives over the True Negatives plus the False Positives. Each point on the curve corresponds to a different classification threshold. The AUC (Area Under the Curve) summarizes the overall performance of the classifier across all thresholds by measuring its ability to distinguish between positive and negative classes. A higher AUC is reflected by the curve (orange line) moving away from the diagonal dotted line (which represents random chance) toward the upper left corner, indicating stronger classification performance.

Figure 4: Impact of SEL Treatment on Engagement Proxies



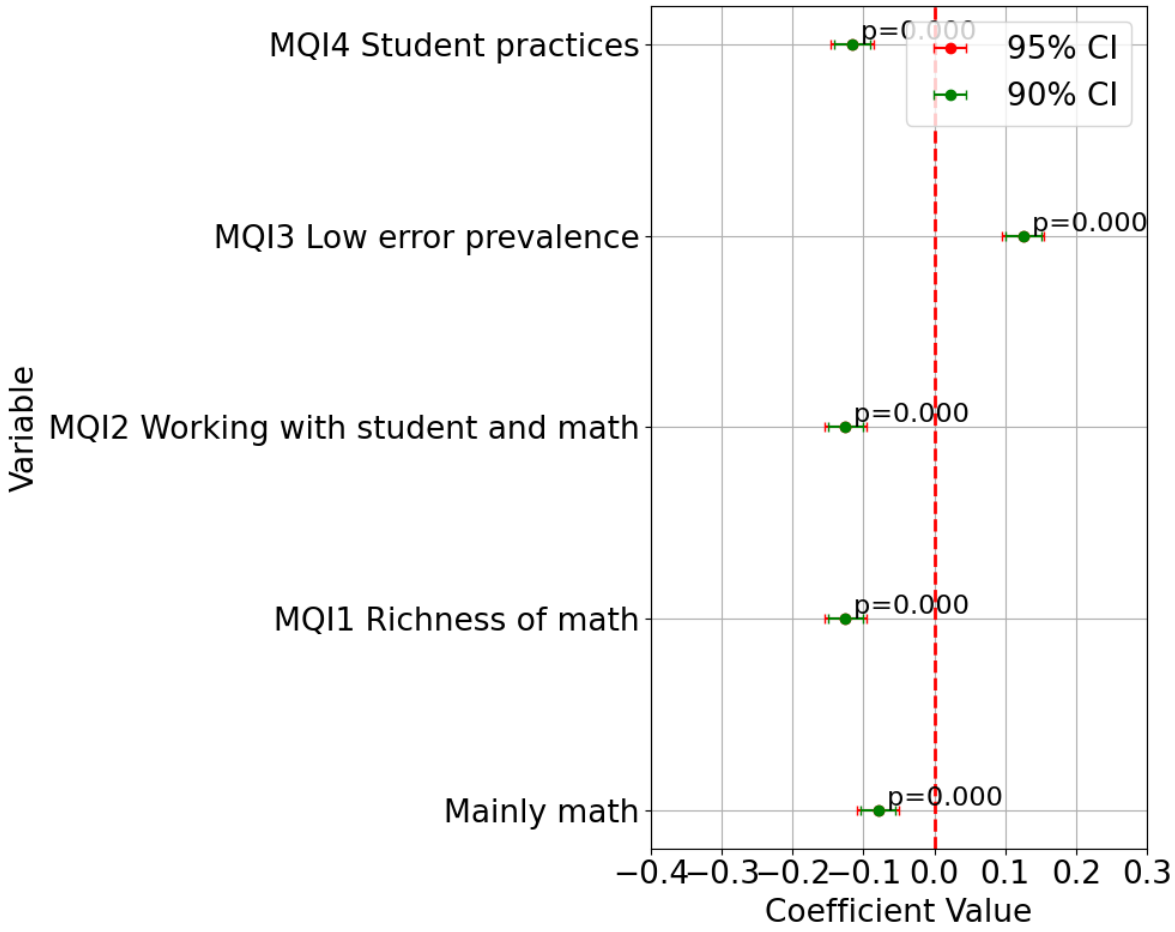
Notes: This figure shows how the text-derived features respond to the SEL treatment (the control group received history classes instead of SEL), estimating Equation 5. Bars represent 95% and 90% confidence intervals.

Figure 5: Impact of SEL on Class Climate



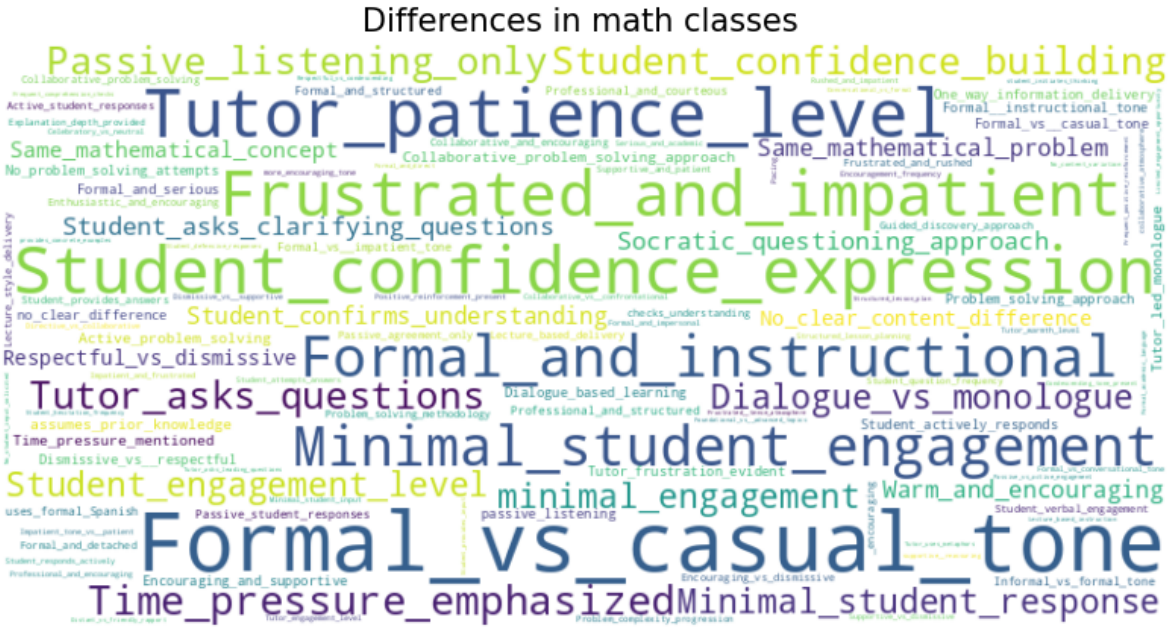
Notes: This figure shows the relationship between the estimated emotions from text using an LLM to the treatment SEL estimating Equation 5. Here the control received history in place of SEL classes. Bars plot 95% and 90% confidence intervals.

Figure 6: Impact of SEL on MQI index



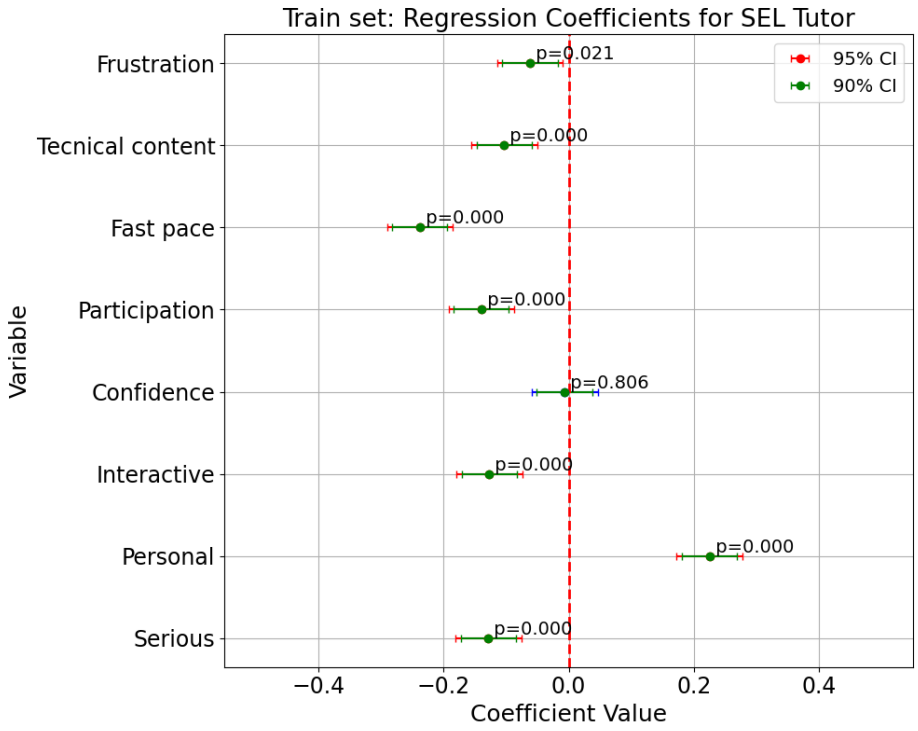
Notes: This figure shows the relationship between the MQI scores coded from text using an LLM to the treatment SEL estimating Equation 5. Here the control received history in place of SEL classes. Bars plot 95% and 90% confidence intervals.

Figure 7: Word Cloud of Differences in High vs Low Prediction Morphs

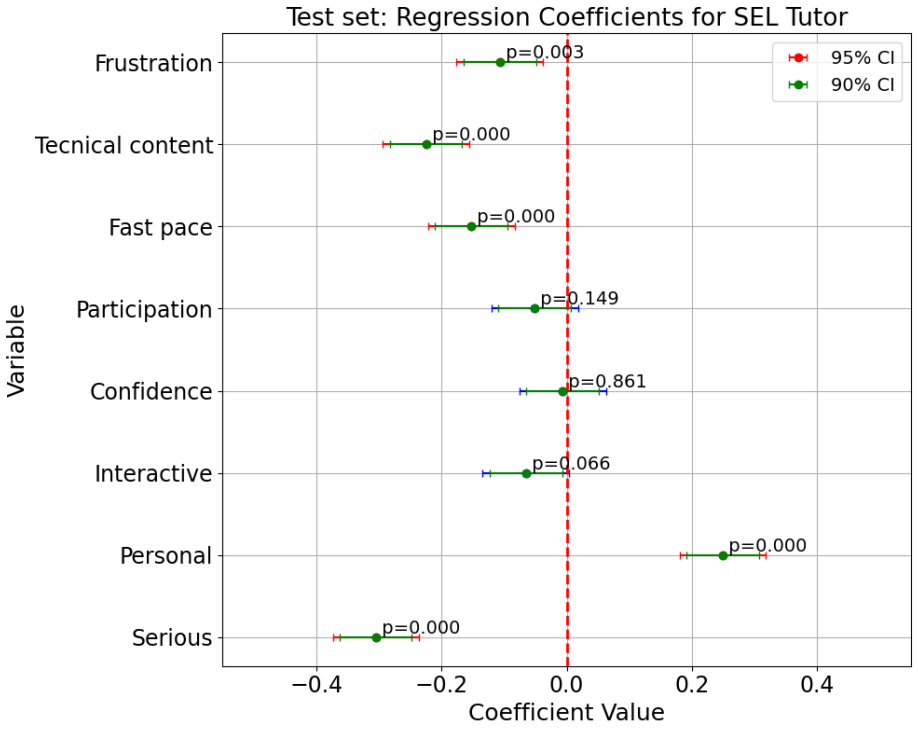


Notes: This figure shows a word cloud of differences between high and low prediction morphs identified by Claude 3.5. To construct the figure, we first created pairwise comparisons between the highest- and lowest-prediction variations and fed them to the Claude 3.5 API, asking it to name the biggest differences. The most frequently mentioned categories across comparisons were then used to rate all samples in the dataset.

Figure 8: Extracted Features



(a) Train set



(b) Test set

Notes: In this figure we show the coefficients resulting from Equation 5. where s_r is the standardized score for each extracted label. As shown in subfigure a) we show this for the train set and in subfigure b) we show the same exercise for the test set for the variables that were relevant in the train set. We plot standardized treatment effects; the bars represent 95% and 90% confidence intervals.

Figure 9: Extracted Features with clustered standard errors

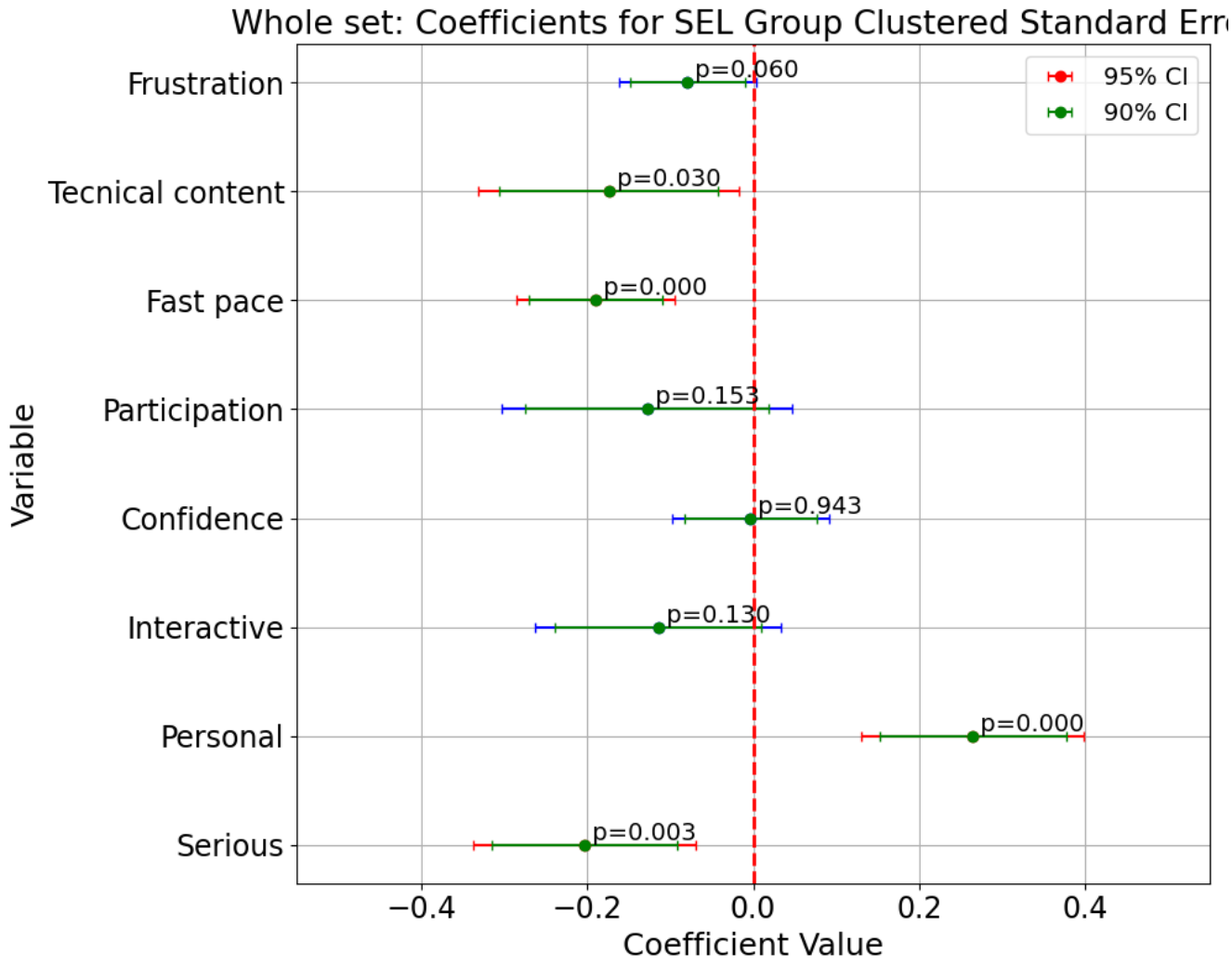


Figure 10: Whole sample

Notes: In this figure we show the coefficients resulting from Equation 5, where s_r is the standardized score for each extracted label. For the whole sample. We plot standardized treatment effects; the standard errors in this regression are clustered at the tutor level, the level of randomization and the bars represent 95% and 90% confidence intervals.

Tables

Table 1: Student's Descriptive Demographics and Balance Tests

Variable	N	(1) 0 Mean/SE	N	(2) 1 Mean/SE	T-test Difference (1)-(2)
Math score baseline	2452	0.018	2559	-0.000	0.019
		(0.020)		(0.020)	
Female Student	2452	0.506	2559	0.506	0.000
		(0.010)		(0.010)	
School grade	2452	8.384	2559	8.335	0.049
		(0.033)		(0.033)	
Income	1442	7600.693	1461	7566.735	33.958
		(175.409)		(169.351)	
Years of education parent	2247	10.961	2341	11.293	-0.332***
		(0.071)		(0.071)	
Parent is single	2284	0.238	2367	0.234	0.005
		(0.009)		(0.009)	
Room to study	2287	0.329	2373	0.343	-0.014*
		(0.010)		(0.010)	
Tutoring	2207	0.097	2302	0.091	0.006
		(0.006)		(0.006)	
Family member went to university	2207	0.608	2302	0.627	-0.019
		(0.010)		(0.010)	
Family member lived in the US	2207	0.081	2301	0.079	0.002
		(0.006)		(0.006)	
Speaks English	2207	0.024	2302	0.028	-0.005
		(0.003)		(0.003)	
Has Wi-FI	2287	0.870	2373	0.871	-0.001
		(0.007)		(0.007)	
Tablet	2287	0.176	2373	0.173	0.003
		(0.008)		(0.008)	
Cellphone	2287	0.845	2373	0.839	0.006
		(0.008)		(0.008)	
Data	2287	0.423	2373	0.402	0.021
		(0.010)		(0.010)	
Electricity	2287	0.973	2373	0.978	-0.005
		(0.003)		(0.003)	
Washing Machine	2287	0.554	2373	0.590	-0.035**
		(0.010)		(0.010)	
TV	2287	0.877	2373	0.878	-0.002
		(0.007)		(0.007)	
Car	2287	0.360	2373	0.388	-0.027**
		(0.010)		(0.010)	
Computer	2287	0.451	2373	0.450	0.001
		(0.010)		(0.010)	
Desk	2287	0.857	2373	0.866	-0.008
		(0.007)		(0.007)	
Attended preschool	2227	0.985	2318	0.988	-0.003
		(0.003)		(0.002)	
Employed	2247	0.503	2339	0.526	-0.023
		(0.011)		(0.010)	

Notes: Fixed effects using variable assignment day are included in all estimation regressions. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level. We present the descriptive demographics of students who participated in the RCT. Column (1) displays the mean statistics for students assigned to the control, while column (2) presents them for students in the SEL treatment. The final column shows the difference between both treatments.The Joint p-value is 0.11.

Table 2: Tutors' Descriptive Demographics and Balance Tests

Variable	N	(1) 0 Mean/SE	N	(2) 1 Mean/SE	T-test Difference (1)-(2)
Family income	364	50054.945	382	49280.105	774.840
		(1736.748)		(1680.816)	
Running water	378	0.989	403	0.988	0.002
		(0.005)		(0.006)	
WiFi	378	0.987	403	0.998	-0.011*
		(0.006)		(0.002)	
WhatsApp	378	0.966	403	0.975	-0.010
		(0.009)		(0.008)	
Electricity	378	0.997	403	0.995	0.002
		(0.003)		(0.004)	
Gas stove	378	0.960	403	0.980	-0.020
		(0.010)		(0.007)	
Washing machine	378	0.865	403	0.893	-0.028
		(0.018)		(0.015)	
Television	378	0.910	403	0.921	-0.011
		(0.015)		(0.013)	
Car	378	0.749	403	0.764	-0.016
		(0.022)		(0.021)	
Age	378	23.175	402	23.199	-0.024
		(0.317)		(0.303)	
Female	378	0.476	402	0.470	0.006
		(0.026)		(0.025)	
GPA	376	88.856	401	89.125	-0.268
		(0.558)		(0.442)	
Private school	378	0.614	402	0.582	0.032
		(0.025)		(0.025)	
Working	378	0.386	402	0.396	-0.009
		(0.025)		(0.024)	
Experience teaching	378	0.439	402	0.458	-0.019
		(0.026)		(0.025)	

Notes: The value displayed for t-tests are the differences in the means across the groups. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.
In this table, we present the descriptive demographics of tutors who participated in the RCT, along with the balance tests across the SEL and history treatments. Column (1) displays the mean statistics for tutors assigned to the control, while column (2) presents the mean statistics for tutors assigned to the SEL treatment. The final column shows the difference between both treatments.

Table 3: Balance Tests for Attrition in Surveys

VARIABLES	(1) Attrition in Math	(2) Attrition in other	(3) Math score baseline	(4) Math score baseline
SEL	0.0158	0.0191	-0.00300	0.00289
	(0.0147)	(0.0147)	(0.0132)	(0.0162)
Observations	5,009	5,009	5,009	3,153
R-squared	0.105	0.107	0.821	0.822
Mean of omitted	.356	.375	.018	.049

Notes: This table reports the results from estimating

$$a_i = \alpha\, SEL_g + \delta_t + \epsilon_{ig}$$

where $a_i$is a dummy variable that takes the value of 1 if student i did not complete the endline survey. We had two endline surveys: a math exam, for which attrition is shown in column (1), and a more comprehensive survey that included our students' tutor attachment survey, for which attrition is shown in column (2). In column (3), we show the balance in initial math scores for all participating students, and in column (4), we present the same balance test conditioned on the students who completed the endline evaluation (where $attrition_math = 0$)

Table 4: Impact of SEL Treatment on Math Scores

	(1)	(2)	(3)
VARIABLES	Math score	Math score	Math score
SEL	0.0936**	0.0848**	0.0744**
	(0.0382)	(0.0357)	(0.0354)
Observations	3,147	3,147	3,124
R-squared	0.097	0.284	0.298
Mean of omitted	-.037	-.037	-.037
Controls		Baseline	Baseline + others

Notes: This table reports the results from estimating Equation 1, where y_i is the standardized score of student i in the endline evaluation. Column (1) shows the specification with no controls except randomization-day fixed effects, Column (2) adds baseline controls, and Column (3) further includes demographic controls such as school grade and gender. The standard error is clustered at the group level.

Table 5: Predictive Contribution of Extracted Labels compared to predicted SEL Scores

	Pred only	Engagement	Class climate	MQI	All	All + pred
Predicted treatment	36.575***					23.289***
	(5.721)					(8.398)
Average time between speakers		-0.019			-0.001	-0.006
		(0.034)			(0.034)	(0.034)
Speaker exchanges		-0.000			-0.000	-0.000
		(0.000)			(0.000)	(0.000)
Number of speakers		0.019**			0.021**	0.020**
		(0.010)			(0.010)	(0.010)
Total silence (seconds)		0.002			0.003	0.003
		(0.002)			(0.002)	(0.002)
Number of questions		-0.000			-0.000	-0.000
		(0.000)			(0.000)	(0.000)
Number of exclamation marks		0.004			0.001	0.002
		(0.002)			(0.002)	(0.002)
Total talk time		0.002			0.003	0.003
		(0.002)			(0.002)	(0.002)
Total class length		-0.002			-0.003	-0.003
		(0.002)			(0.002)	(0.002)
Not neutral			-1.178		-1.370*	-1.448*
			(0.835)		(0.817)	(0.815)
Not frustrated			-1.328		-1.482*	-1.572*
			(0.845)		(0.821)	(0.820)
Satisfied			1.289		1.590*	1.649**
			(0.843)		(0.827)	(0.824)
Rapport			-0.015		0.084	0.090
			(0.064)		(0.067)	(0.067)
MQI1 Richness of math				-0.074	-0.001	0.033
				(0.068)	(0.072)	(0.073)
MQI2 Working with student				0.070	-0.005	0.043
				(0.069)	(0.075)	(0.077)
MQI3 Low error prevalence				0.078**	-0.015	-0.024
				(0.040)	(0.047)	(0.047)
MQI4 Student practices				-0.228**	-0.336***	-0.360***
				(0.095)	(0.097)	(0.097)
Mainly math				0.062	0.092	0.085
				(0.069)	(0.069)	(0.069)
R-squared	0.036	0.033	0.007	0.037	0.083	0.090
R-squared Adj.	0.035	0.026	0.003	0.032	0.069	0.075
Observations	1102	1102	1102	1102	1102	1102

Standard errors in parentheses $*p < .1, **p < .05, ***p < .01$

Notes: Validation of the contribution of extracted labels to predicting SEL scores. We estimate Equation 3. The table reports the coefficient of the predictor, standard errors, number of observations, R^2, adjusted R^2, and notes on which variables are included in each specification. Each model demonstrates the relative contribution of the predictor, and known features to explaining the variation in SEL outcomes.

Table 6: Contribution of Extracted Labels to predicted SEL Scores

	Engagement	Class climate	MQI	Labels	All	All + Labels
Avg time between speakers	0.000				0.000*	0.000**
	(0.000)				(0.000)	(0.000)
Speaker exchanges	-0.000				-0.000**	-0.000*
	(0.000)				(0.000)	(0.000)
Number of speakers	0.000***				0.000	0.000
	(0.000)				(0.000)	(0.000)
Total silence (seconds)	0.000				0.000	0.000*
	(0.000)				(0.000)	(0.000)
Number of questions	-0.000***				-0.000***	-0.000***
	(0.000)				(0.000)	(0.000)
Number of exclamation marks	-0.000*				-0.000***	-0.000***
	(0.000)				(0.000)	(0.000)
Total talk time	0.000				0.000	0.000*
	(0.000)				(0.000)	(0.000)
Total class length	-0.000				-0.000	-0.000*
	(0.000)				(0.000)	(0.000)
Not neutral		0.008*			0.003	0.003
		(0.004)			(0.003)	(0.002)
Not frustrated		0.007*			0.004	0.004
		(0.004)			(0.003)	(0.003)
Satisfied		-0.008*			-0.003	-0.003
		(0.004)			(0.003)	(0.003)
Rapport		-0.001***			-0.000	-0.000
		(0.000)			(0.000)	(0.000)
MQI1 Richness of math			-0.002***		-0.001***	-0.000
			(0.000)		(0.000)	(0.000)
MQI2 Working with student h			-0.002***		-0.002***	-0.001***
			(0.000)		(0.000)	(0.000)
MQI3 Low error prevalence			0.000**		0.000**	0.000
			(0.000)		(0.000)	(0.000)
MQI4 Student practices			0.001***		0.001***	0.000
			(0.000)		(0.000)	(0.000)
Mainly math			0.000		0.000	-0.000
			(0.000)		(0.000)	(0.000)
Serious				-0.001***		-0.000***
				(0.000)		(0.000)
personal				-0.001***		-0.001***
				(0.000)		(0.000)
Fast pace				-0.001***		-0.001***
				(0.000)		(0.000)
Technical content				-0.001***		-0.001***
				(0.000)		(0.000)
Frustration				-0.000		-0.000***
				(0.000)		(0.000)
R-squared	0.291	0.063	0.424	0.512	0.555	0.683
R-squared Adj.	0.286	0.060	0.421	0.510	0.548	0.676
Observations	1102	1102	1102	1102	1102	1102

Standard errors in parentheses $*p < .1, **p < .05, ***p < .01$

Notes: Validation of the contribution of extracted labels to predicting SEL scores. The table reports the coefficient of the predictor, standard errors, number of observations, R^2, adjusted R^2, and notes on which variables are included in each specification. Each model demonstrates the relative contribution of the predictor, extracted labels, and text features to explaining variation in SEL outcomes.

Table 7: Example Transcripts and Variations (Morphs)

Original	**Highest**	**Lowest**
T: Hello Student1, good afternoon.	T: Hello Student1, how are you?	T: Hello Student1, how are you?
S: Good afternoon, teacher.	S: Good afternoon, teacher.	S: Fine, thanks.
T: How are you?	T: How are you?	T: Excellent. Recently you've been
S: Good.	S: Good.	covering linear equations,
T: Good?	T: Are you sure?	fundamental to algebra. These equa-
S: Yes.	S: Yes.	tions are characterized by...
T: How were you doing these days of	T: How did you do in class?	...
classes?	S: Good.	...
S: Also good.	T: Any new topics?	... Do you have any questions?
T: Any new topics?	S: No.	
S: No.	T: What did you see?	
T: What did you see these days?	S: Linear equations.	
S: Simply a little about linear equa-		
tions.		

Notes: Example transcripts: Original, Highest SEL probability, and Lowest SEL probability variations.

Table 8: Example Transcripts in Labeled Dataset

Text	**Treated**	**Simple language**
Well, let's get started. And student1. And yesterday we left off with this exercise that gave us minus 41 over eighteen. We are doing quite complex operations now, already quite long, with many fractions ... possible synchronization. Whatever problem they put in front of you, I think now you can solve it without any trouble.	0	4
Can you hear me? Hello, teacher. How are you? Fine. That's good! Sorry, I couldn't get into the session. I was having problems, that's why I was late. No problem, teacher. Thanks. Are you ready to start? Yes!	0	6
Hi, how are you? Well, I feel a little sick in my throat, but I'm okay. Oh, you got sick, right? Last class it did sound like you weren't feeling well, but the good thing is that you're already recovering. Have you been to school lately? No, I didn't go this week, I'll go next week. That's good, it's good that you took care of yourself and that you're getting over your illness. It was just a sore throat, nothing serious.	1	6
⋮	⋮	⋮

Notes: Example transcripts: Labelling dataset after generated hypotheses.

Table 9: Predictive Contribution of Extracted Labels to SEL Scores

	Pred Only	Labels	Pred + labels	All controls
Predicted treatment	36.575***		44.665***	
	(5.721)		(8.047)	
Serious		-0.103***	-0.077***	-0.099***
		(0.028)	(0.028)	(0.030)
personal		0.022	0.068***	0.031
		(0.023)	(0.024)	(0.026)
Fast pace		-0.001	0.052**	0.008
		(0.024)	(0.026)	(0.025)
Technical content		-0.002	0.051**	0.048*
		(0.022)	(0.023)	(0.028)
Frustration		-0.031	-0.026	-0.007
		(0.021)	(0.021)	(0.022)
Average time between speakers				0.003
				(0.033)
Speaker exchanges				-0.000
				(0.000)
Number of speakers				0.019**
				(0.010)
Total silence (seconds)				0.003
				(0.002)
Number of questions				-0.000
				(0.000)
Number of exclamation marks				-0.000
				(0.002)
Total talk time				0.003
				(0.002)
Total class length				-0.003
				(0.002)
Not neutral				-1.437*
				(0.810)
Not frustrated				-1.556*
				(0.815)
Satisfied				1.622**
				(0.819)
Rapport				0.059
				(0.070)
MQI1 Richness of math				0.007
				(0.083)
MQI2 Working with student and math				0.058
				(0.077)
MQI3 Low error prevalence				-0.003
				(0.048)
MQI4 Student practices				-0.358***
				(0.097)
Mainly math				0.136**
				(0.069)
R-squared	0.036	0.047	0.073	0.105
R-squared Adj.	0.035	0.043	0.068	0.087
Observations	1102	1102	1102	1102

Standard errors in parentheses. $*p < .1, **p < .05, ***p < .01$

Notes: Validation of the contribution of extracted labels to predicting SEL scores. We estimate Equation 3 where L_r represents the extracted label variables (shown in Figure A.III). The table reports the coefficient of the predictor, standard errors, number of observations, R^2, adjusted R^2, and notes on which variables are included in each specification. Each model demonstrates the relative contribution of the predictor, extracted labels, and text features to explaining variation in SEL outcomes.

References

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.

Ang, Rebecca P, Soo Lin Ong, and Xiang Li, "Student version of the Teacher–Student Relationship Inventory (S-TSRI): Development, validation and invariance," *Frontiers in Psychology*, 2020, *11*, 519204.

Angrist, Noam, Peter Bergman, and Moitshepi Matsheng, "School's Out: Experimental Evidence on Limiting Learning Loss Using "Low-Tech" in a Pandemic," Technical Report, National Bureau of Economic Research 2020.

Bain, Max, Jaesung Huh, Tengda Han, and Andrew Zisserman, "WhisperX: Time-Accurate Speech Transcription of Long-Form Audio," 2023.

Carlana, Michela and Eliana La Ferrara, "Apart But Connected: Online Tutoring and Student Outcomes During the COVID-19 Pandemic," CEPR Discussion Paper DP15761, Centre for Economic Policy Research February 2021. Available at SSRN: `https://ssrn.com/abstract=3783987`.

Center for Education Policy Research, Harvard University, "Mathematical Quality of Instruction (MQI) Rubric," n.d. Observational rubric.

Chioda, Laura, David Contreras-Loya, Paul Gertler, and Dana Carney, "Making entrepreneurs: Returns to training youth in hard versus soft business skills," Technical Report, National Bureau of Economic Research 2021.

Cipriano, Christina, Michael J Strambler, Lauren H Naples, Cheyeon Ha, Megan Kirk, Miranda Wood, Kaveri Sehgal, Almut K Zieher, Abigail Eveleigh, Michael McCarthy et al., "The state of evidence for social and emotional learning: A contemporary meta-analysis of universal school-based SEL interventions," *Child development*, 2023, *94* (5), 1181–1204.

_ , Tia N Barnes, Susan E Rivers, and Marc Brackett, "Exploring changes in student engagement through the ruler approach: An examination of students at risk of academic failure," *Journal of Education for Students Placed at Risk (JESPAR)*, 2019, *24* (1), 1–19.

Collaborative for Academic, Social, and Emotional Learning (CASEL), "CASEL's SEL Framework: What Are the Core Competence Areas and Where Are They Promoted?," `https://casel.org/casel-sel-framework-11-2020/` October 2020. Accessed: 2025-07-08.

Corcoran, Roisin P, Alan CK Cheung, Elizabeth Kim, and Chen Xie, "Effective universal school-based social and emotional learning programs for improving academic achievement: A systematic review and meta-analysis of 50 years of research," *Educational Research Review*, 2018, *25*, 56–72.

Cornelius-White, Jeffrey, "Learner-centered teacher-student relationships are effective: A meta-analysis," *Review of educational research*, 2007, *77* (1), 113–143.

Deming, David J, "The growing importance of social skills in the labor market," *The quarterly journal of economics*, 2017, *132* (4), 1593–1640.

Demszky, Dora, Jing Liu, Zid Mancenido, Julie Cohen, Heather C. Hill, Dan Jurafsky, and Tatsunori Hashimoto, "Measuring Conversational Uptake: A Case Study on Student–Teacher Interactions," *Annenberg Institute for School Reform at Brown University*, 2021, (21-424). ERIC Number: ED613663.

Demszky, Dorottya, Jing Liu, Heather C Hill, Shyamoli Sanghi, and Ariel Chung, "Improving Teachers' Questioning Quality through Automated Feedback: A Mixed-Methods Randomized Controlled Trial in Brick-and-Mortar Classrooms. EdWorkingPaper No. 23-875.," *Annenberg Institute for School Reform at Brown University*, 2023.

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.

Falk, Armin, Fabian Kosse, and Pia Pinger, "Mentoring and schooling decisions: Causal evidence," Technical Report 13387, Institute of Labor Economics (IZA) Discussion Papers 2020.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy, "Text as data," *Journal of Economic Literature*, 2019, *57* (3), 535–574.

Gertler, Paul, James Heckman, Rodrigo Pinto, Arianna Zanolini, Christel Vermeersch, Susan Walker, Susan M Chang, and Sally Grantham-McGregor, "Labor market returns to an early childhood stimulation intervention in Jamaica," *Science*, 2014, *344* (6187), 998–1001.

Hagelskamp, Carolin, Marc A Brackett, Susan E Rivers, and Peter Salovey, "Improving classroom quality with the RULER approach to social and emotional learning: Proximal and distal outcomes," *American Journal of Community Psychology*, 2013, *51*, 530–543.

Heckman, James J and Tim Kautz, "Hard evidence on soft skills," *Labour economics*, 2012, *19* (4), 451–464.

Hoffmann, Jessica D, Marc A Brackett, Craig S Bailey, and Cynthia J Willner, "Teaching emotion regulation in schools: Translating research into practice with the RULER approach to social and emotional learning.," *Emotion*, 2020, *20* (1), 105.

Hugging Face, "FacebookAI/xlm-roberta-base model card," `https://huggingface.co/FacebookAI/xlm-roberta-base` 2022. Accessed: 2025-08-08.

Jr, Roland G Fryer, "Information, non-financial incentives, and student achievement: Evidence from a text messaging experiment," *Journal of Public Economics*, 2016, *144*, 109–121.

Kane, Thomas J, Daniel F McCaffrey, Trey Miller, and Douglas O Staiger, "Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment. Research Paper. MET Project.," *Bill & Melinda Gates Foundation*, 2013.

Kim, Gwanghee, Ick Hoon Jin, and Minjeong Jeon, "Analyzing Classroom Interaction Data Using Prompt Engineering and Network Analysis," *arXiv preprint arXiv:2501.18912*, 2025.

Kraft, Matthew A, "Interpreting effect sizes of education interventions," *Educational researcher*, 2020, *49* (4), 241–253.

Llanes, Salome Aguilar, Bernardo Garcia Bulle Bueno, Maria Fernanda Albo Alarcón, and Sabastián Guevara Cota, "Expanding Access to Tutoring: A Scalable Platform for Personalized Learning and Data-Driven Research," in "International Conference on Artificial Intelligence in Education" Springer 2025, pp. 364–371.

Ludwig, Jens and Sendhil Mullainathan, "Machine learning as a tool for hypothesis generation," *The Quarterly Journal of Economics*, 2024, *139* (2), 751–827.

Merriam, Sharan B and Lisa M Baumgartner, *Learning in adulthood: A comprehensive guide*, John Wiley & Sons, 2020.

Movva, Rajiv, Kenny Peng, Nikhil Garg, Jon Kleinberg, and Emma Pierson, "Sparse autoencoders for hypothesis generation," *arXiv preprint arXiv:2502.04382*, 2025.

Nathanson, Lori, Susan E Rivers, Lisa M Flynn, and Marc A Brackett, "Creating emotionally intelligent schools with RULER," *Emotion review*, 2016, *8* (4), 305–310.

Nickow, Andre, Philip Oreopoulos, and Vincent Quan, "The Impressive Effects of Tutoring on PreK-12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence," Working Paper, National Bureau of Economic Research 2020.

OECD, "OECD Survey on Social and Emotional Skills 2023 Technical Report," Technical Report, OECD Publishing, Paris May 2023.

Perez, Matthew, Berk Coker, Kemal Berk Kocabagli, Jessica Vitale, and Alyssa Van Camp, "Multimodal Classroom Diarization with GPT Re-scoring: Teacher or Student?," in "International Conference on Artificial Intelligence in Education" Springer 2025, pp. 228–235.

Pianta, Robert C., Karen M. La Paro, and Bridget K. Hamre, *Classroom Assessment Scoring System (CLASS) Manual, Pre-K*, Baltimore, MD: Paul H. Brookes Publishing, 2008.

Resnjanskij, Sven, Jens Ruhose, Simon Wiederhold, Ludger Woessmann, and Katharina Wedel, "Can Mentoring Alleviate Family Disadvantage in Adolescence? A Field Experiment to Improve Labor Market Prospects," *Journal of Political Economy*, 2024, *132* (3), 1013–1062.

Sawaya, Sandra, Jennifer Jacobs, Robert Moulder, Chelsea Chandler, Brent Milne, Tom Fischaber, and Sidney K D'Mello, "Improving Tutor Discourse Practices via AI-Enhanced Coaching: A Piecewise Latent Growth Curve Modeling Approach," in "International Conference on Artificial Intelligence in Education" Springer 2025, pp. 76–89.

Schunk, Daniel, Eva M Berger, Henning Hermes, Kirsten Winkel, and Ernst Fehr, "Teaching self-regulation," *Nature human behaviour*, 2022, *6* (12), 1680–1690.

Skoog, Alexandra, Asher A. Miller, Rista C. Plate, Duncan C. Meyers, Andrew S. Tucker, Gabrielle Meyers, Melissa Kay Diliberti, Heather L. Schwartz, Megan Kuhfeld, Robert J. Jagers et al., "Social and Emotional Learning in US Schools," *RAND Corporation*, 2022. Available at `https://www.rand.org/pubs/research_reports/RRA397-3.html`.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, 2017, *30*.

Vitale, Jessica, Kemal Berk Kocabagli, Shyamoli Sanghi, Alyssa Van Camp, Sarah Miller, and Berk Coker, "From Noisy Classroom Transcripts to Actionable Feedback: Fine-Tuning GPT-4o to Detect Teachers' Opportunities to Respond," in "International Conference on Artificial Intelligence in Education" Springer 2025, pp. 149–162.

Walton, Kate E, Jeremy Burrus, Cristina Anguiano-Carrasco, Jason Way, and Dana Murano, "Aligning mosaic by ACT: Social emotional learning to the collaborative for academic, social, and emotional learning (CASEL) Framework," 2021.

Walton, Kate, Jeff Allen, Maxwell J Box, Dana Murano, and Jeremy Burrus, "Social and Emotional Skills Predict Postsecondary Enrollment and Retention," *Journal of Intelligence*, 2023, *11* (10), 186.

Wang, Rose E., Ana T. Ribeiro, Carly D. Robinson, Susanna Loeb, and Dora Demszky, "Tutor CoPilot: A Human-AI Approach for Scaling Real-Time Expertise," 2025.

Weidmann, Ben and David J Deming, "Team players: How social skills improve team performance," *Econometrica*, 2021, *89* (6), 2637–2657.

Zhang, Qingyang, Rose E. Wang, Ana T. Ribeiro, Dora Demszky, and Susanna Loeb, "Educator Attention: How Computational Tools Can Systematically Identify the Distribution of a Key Resource for Students," 2025.

A Appendix

A.1 SEL activities

This is a detailed description of the content covered in each of the SEL modules:

- Activity 1: How Do I Feel? The main activity of the class revolves around students answering three questions: What emotions do I know? How do I feel? And, what emotions have I experienced recently? This is followed by drawing a picture of a time they felt a certain way they chose to express.

- Activity 2: My Strengths and Opportunities. In this session, tutors guide students to recognize and share their strengths and opportunities as individuals, in order to create an action plan and continue learning from them. They reflect on what they want to improve, recognize what they can do to achieve it, identify where they can start, value what they have achieved and what is missing, and plan the next steps.

- Activity 3: Pause to Breathe. The objective is to teach students how to regulate their emotions using breathing exercises. During the class, students share a moment they found uncomfortable or difficult, are asked how they felt, how they acted, and what the consequence was. After sharing, they are shown the five-point breathing technique and anchor breathing.

- Activity 4: Gratitude Feels Good. This session begins with applying the anchor breathing technique learned in the previous class. The tutor then asks what good situations or experiences the students have had recently and how it feels to receive support, care, and affection from others. They mark their feelings on an emotion chart, recognize whom they are grateful to for the positive things in their lives, and conclude by writing a thank-you note or drawing for the person, expressing how they feel receiving their support and help.

- Activity 5: What Can I Do to Regain Calm? This lesson is divided into two activities: a reflective one and a learning one. First, the class shares what emotions they have experienced during the week, how they reacted when angry, scared, or sad, the consequences of reacting in aggressive, impulsive, or evasive ways, and how they can handle such situations differently to regain calm. After the discussion, the tutor shares behaviors they can adopt in such situations, such as talking to someone they trust, taking time alone to calm down, deep breathing, going for a walk, listening to pleasant music, and writing or drawing about their feelings. It ends with asking them to draw the strategy they would use based on what they learned and how they would feel applying it.

- Activity 6: What Do You Like About Yourself? This session focuses on self-awareness and self-esteem, prompting students to question which aspect of their personality they like the most, why they like it, and recognizing how it contributes to their well-being and helps them overcome daily challenges. They are then invited to write a letter thanking themselves for this aspect of their personality.

- Activity 7: What I Like and What I Don't Like. In this segment on autonomy and self-awareness, students are asked to divide a sheet of paper into four equal parts. In each quadrant, they write the following: describe something they dislike a lot and how it makes them feel, describe something they dislike a little and how it makes them feel, describe something they like and how it makes them feel, and describe something they like a lot and how it makes them feel. After answering the questions in each designated space, they reflect together on why it is important to recognize what they like and dislike the most.

- Activity 8: Learning from Mistakes. In this final module on growth mindset, they discuss how the mistakes they make in daily life can sometimes affect how they feel about themselves. Everyone shares if they have made a mistake that has affected them emotionally, and once these ideas are grounded, they are asked to draw a tree where at the base of the trunk they write the mistake they identified, on the trunk they write the emotions it caused, and on the branches of the tree the aspects of their lives that were affected by the action. Finally, they are invited to recognize what they learned from that experience.

A.2 Recordings' data

Appendix Table A.I: Summary of recordings

Metric	Count
Assigned classes	37,369
Classes taught	30,081
Classes attended	25,960
Classes with file	24,784
Classes with processed transcript	20,486
Math classes with processed transcript	11,731

Notes: This table summarizes the progression from the total number of classes assigned to tutors in the program to the final analytic sample used in the empirical analysis. Starting from all classes assigned to tutoring groups over the study period, we state the number of classes reported in the platform by tutors as classes taught, the number of those classes attended by students, and among those, the number for which at least one file with the recording was uploaded. Some uploaded files are corrupted or in an unsupported format; we therefore report separately the number of classes with successfully processed files for which a transcript was generated through the automated transcription pipeline. The final row shows the number of math-only classes, which constitute the analytic sample used throughout the analysis.

Appendix Table A.II: Example of Data Used to Train Predictor

Spanish:

{label: 1, text: hola *estudiante1*. hola. ¿como estas? bien, ¿y usted? bien. ¿te acuerdas que estabamos haciendo la vez pasada? vimos numeros positivos y negativos y fracciones. ah, cierto, cierto, cierto. aja, ok. pero... ¿acabamos eso, no? creo que si. ok. pues entonces te pregunto, ¿que quieres revisar hoy, ya que llegaste primero? pues, no tengo nada de apreciacion. no tienes... ¿no tienes como? que no tengo dudas. ah, ok. entonces, que sera bueno. ¿que estas viendo ahorita?}

English translation:

{label: 1, text: hello *student1*. hello. how are you? good, and you? good. do you remember what we were doing last time? we looked at positive and negative numbers and fractions. ah, right, right, right. uh-huh, ok. but... we finished that, didn't we? I think so. ok. so then let me ask you, what do you want to review today, since you arrived first? well, I don't have anything in mind. you don't... you don't have what? that I don't have any doubts. ah, ok. so then, what would be good. what are you learning right now?}

Spanish:

{label: 0, text: ok, ya envie la liga a whatsapp. ninguno lo ha leido, pero estoy esperando que se conecten los alumnos. hola, buenas tardes. ¿como estas, *estudiante1*? muy bien, gracias. pues, mira, ¿que sera? envienla a las cinco, ya son como cinco, ocho, entonces, yo creo que ya comenzamos, y si cualquiera se une, pues, entonces, no, no estariamos ayudando en el tema. muy bien. ¿como te sientes como con ganas de ver matematicas? porque hoy nos tocan matematicas. si, perfecto. pues mira, en la tutorial pasada terminamos de ver lo que era la proporcionalidad inversa y ya cerramos ese tema. ok, y como me habias comentado, querias ver el tema de las ecuaciones, el sistema de ecuaciones y sus soluciones de que eran de dos por dos. bueno, que estos dos por dos de ecuaciones. ok, entonces es lo que vamos a revisar el dia de hoy. vamos a comenzar con el tema.}

English translation:

{label: 0, text: ok, I already sent the link on WhatsApp. none of them have read it, but I'm waiting for the students to connect. hello, good afternoon. how are you, *student1*? very good, thank you. well, look, what would it be? send it at five, it's already about five-oh-eight, so I think we should start, and if anyone joins, well, then, no, no, we wouldn't be holding them back on the topic. very good. how are you feeling—ready to do some math? because today we have math. yes, perfect. well look, in the last tutorial we finished looking at what inverse proportionality was and we closed that topic. ok, and as you had mentioned, you wanted to look at the topic of equations, the system of equations and their solutions when they were two-by-two. well, those two-by-two systems of equations. ok, so that's what we're going to review today. let's start with the topic.}

A.3 Attendance

Appendix Table A.III: Impact of SEL Treatment on Number of Classes and Students' Attendance

	(1)	(2)
VARIABLES	Number of classes	Attendance
SEL	-0.431	-0.343
	(0.488)	(0.324)
Observations	1,140	4,991
R-squared	0.505	0.222
Mean of omitted	25.497	10.528

Notes: This table reports the results from estimating Equation 1 where y_i is the average attendance of student i or the average number of classes taught by the tutor in a specific group. Column (1) shows the number of classes taught by tutors in the treated groups, with the regression at the group level. Column (2) shows the average attendance for students in the control group. The standard error is clustered at the group level.

A.4 Tracking

Appendix Table A.IV: Impact of Tracking and SEL Intervention on Math Scores

	(1)	(2)	(3)
VARIABLES	Math score	Math score	Math score
Only SEL	0.103***	0.0910***	0.0835**
	(0.0367)	(0.0347)	(0.0345)
Tracking	0.0275	0.0451	0.0403
	(0.0368)	(0.0333)	(0.0332)
Tracking and SEL	0.0563	0.0535	0.0501
	(0.0363)	(0.0355)	(0.0351)
Observations	6,412	6,407	6,373
R-squared	0.060	0.243	0.254
Mean of omitted	-.033	-.033	-.033
Controls		Baseline	Baseline + others

Notes: In this table, we present the results of the regression

$$y_i = \alpha_0 + \alpha_1 T_g + \alpha_2 SEL_g + \alpha_3 T_SEL_g + X_i\beta + \delta_t + \epsilon_{ig}$$

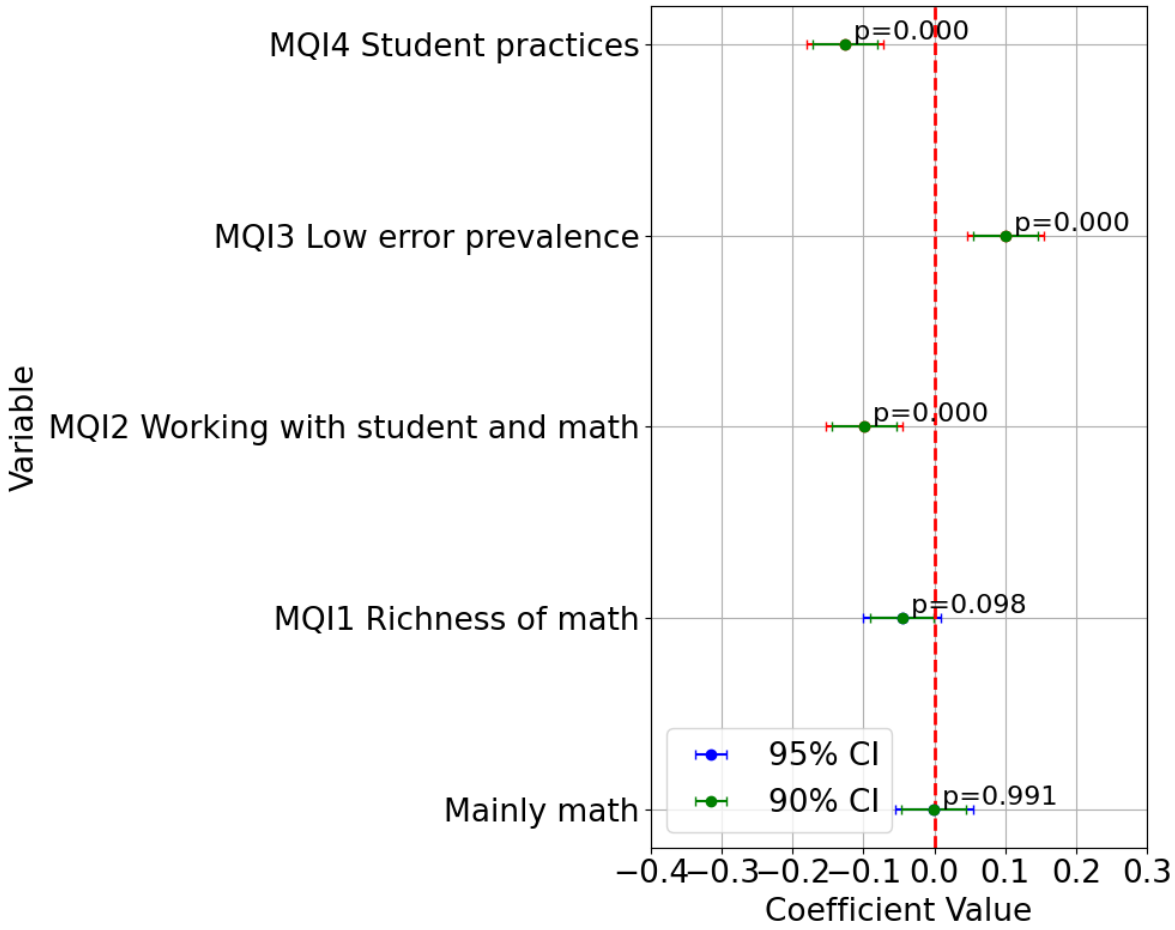
where y_i represents the standardized math score of student i, T_g indicates whether group g was tracked and did not receive SEL content classes, SEL_g indicates whether group g received SEL content and was not tracked and T_SEL_g indicates that group g was both tracked and received SEL classes. The vector X_i comprises controls for student i, including baseline evaluations and other demographic variables. We include fixed effects δ_t for randomization day t, since the intervention was implemented in waves across multiple dates. Column (1) shows the specification with no controls except randomization-day fixed effects. Column (2) adds baseline controls, and Column (3) further includes demographic controls such as school grade and gender. The results presented are the pooled results comparing the SEL and History treatment arms, including both tracked and non-tracked groups. The standard error is clustered at the group level.

As shown in Table A.IV, the treatment effect appears to be larger among groups that were not tracked, which could suggest that SEL are more effective in more diverse group settings. However, the difference between SEL tracked and non-tracked groups is not statistically significant, nor is the difference between tracked and non-tracked history, which is why we pool the groups in our main specification. If anything, this implies that our estimated SEL effects are conservative.

A.5 MQI

Mathematics quality of instruction (MQI).

Appendix Figure A.I: Impact of SEL on Dimensions of Mathematics Quality of Instruction (MQI)



Notes: This figure shows how each rated dimension of the Mathematics Quality of Instruction (MQI) index, extracted from minutes 12 to 24 of classroom transcripts with a large-language model, responds to the SEL treatment (the control group received history lessons instead of SEL) estimating Equation 5. We plot standardized treatment effects for the overall index; bars represent 95% and 90% confidence intervals.

A.6 Snippet level tables

Appendix Table A.V: Predictive Contribution of Known Variables to SEL Scores

	Pred Only	Labels	Class climate	MQI	All
Predicted treatment	26.551***				
	(2.863)				
Serious		-0.056***			-0.053***
		(0.013)			(0.013)
personal		0.011			0.003
		(0.010)			(0.011)
Fast pace		0.003			0.013
		(0.011)			(0.012)
Technical content		-0.011			0.000
		(0.010)			(0.013)
Frustration		-0.018**			-0.013
		(0.009)			(0.010)
Not neutral			-0.488		-0.548
			(0.355)		(0.352)
Not frustrated			-0.516		-0.588*
			(0.359)		(0.354)
Satisfied			0.539		0.674*
			(0.358)		(0.355)
Rapport			-0.003		-0.007
			(0.028)		(0.030)
MQI1 Richness of math				-0.039	0.018
				(0.030)	(0.035)
MQI2 Working with student and math				0.001	-0.031
				(0.030)	(0.032)
MQI3 Low error prevalence				0.059***	0.018
				(0.018)	(0.021)
MQI4 Student practices				-0.102***	-0.167***
				(0.039)	(0.041)
Mainly math				0.048*	0.085***
				(0.028)	(0.028)
R-squared	0.027	0.026	0.003	0.024	0.045
R-squared Adj.	0.027	0.024	0.001	0.022	0.041
Observations	3098	3098	3098	3098	3098

Standard errors in parentheses.
$*p < .1, **p < .05, ***p < .01$

Notes: Validation of the contribution of extracted labels to predicting SEL scores. We estimate Equation 3 in the hold-out set sample. The table reports the coefficient of the predictor, standard errors, number of observations, R^2, adjusted R^2, and notes on which variables are included in each specification. Each model demonstrates the relative contribution of the predictor, and known features to explaining the variation in SEL outcomes.

A.7 Prompts

Appendix Table A.VI: Refined Prompt to Generate Variations Based on Existing Interactions

You are a writer. Your task is to take an original online tutoring interaction between a tutor and a student and generate multiple rewritten versions of that same interaction.

Each version must: - Maintain the central theme and learning objective of the original conversation.

- Change the tone, difficulty, style, or dynamics of the interaction according to the instructions provided.

- Sound natural and coherent, like a real conversation between a tutor and a student in an online environment.

- Maintain the dialogue format (alternating turns between tutor and student).

- Not include external narration, notes, or explanations (do not write "Scene: ...", "Description: ...", etc.).

- Use fluent and natural language (in the same language as the original conversation).

- Reflect the digital context, if applicable (e.g., "share your screen," "look at the chart," "type the response in the chat").

You will receive a set of fields with descriptions, for example: 'versionX: str = Field(description="Rewrite the original interaction by making...")' For each description, you must produce a complete version of the conversation that: - Clearly expresses the indicated change.

- Is noticeably different from the original conversation and the other versions.

- Maintains a length, pace, and consistency similar to a real online tutorial.

Output Format: Returns a JSON object with the same field names ('version1', 'version2', ... 'version30'), where each field contains the complete text of the rewritten conversation.

The categories are as follows, rewrite the original interaction so that:

* the tutor is impatient, rude, and intolerant of the student.

* the tutor is extremely patient, kind, and empathetic.

* the tutor directly provides the answers without encouraging the student's reasoning.

* the tutor uses a Socratic style, guiding with questions instead of giving answers.

* the student is shy, insecure, and speaks very little.

* the student actively participates, speaks a lot, and asks many questions.

* the tutor speaks using technical, formal language and in long monologues.

* the environment is pleasant, collaborative, and fosters a good relationship between tutor and student.

* the conversation is much shorter than the original, while maintaining the same central theme. * Rewrite the original interaction by changing the outcome of the conversation.

* Rewrite the original interaction by having the tutor use humor, jokes, or cultural references while explaining.

* Rewrite the original interaction by having the tutor use digital tools or screen references (e.g., "share your screen," "look at the graph," etc.) typical of online tutoring.

* the student becomes frustrated or argues with the tutor.

* the tutor uses analogies or everyday examples to explain the concepts.

* the tutor speaks very little and lets the student construct almost all the reasoning.

* both characters stray from the main topic with anecdotes or irrelevant comments.

* the student is distracted or multitasking (for example, responding slowly or with confusion).

* the tutor is enthusiastic, motivating, and positively reinforces each of the student's responses.

* the conversation takes place in an evaluative or oral exam tone, with tension and pressure.

* the student has great difficulty understanding the topic, and the tutor has to constantly simplify the language. * Rewrite the original interaction so the student is very skilled and understands quickly, allowing the tutor to move on to more complex topics.

* using more colloquial language, with informal expressions typical of online tutoring among young people.

* using more technical and formal language, as if the tutor were coaching an advanced student.

* so the student frequently loses focus, requiring the tutor to redirect them several times.

* so the tutor asks constant verification questions, ensuring the student understands each step.

* by increasing the difficulty of the exercises and the tutor's mathematical explanations.

* by reducing the difficulty of the math.

Notes: This table shows a refined prompt used with Claude 3.5 to extract differences between pairwise comparisons of *morphed* phrases.

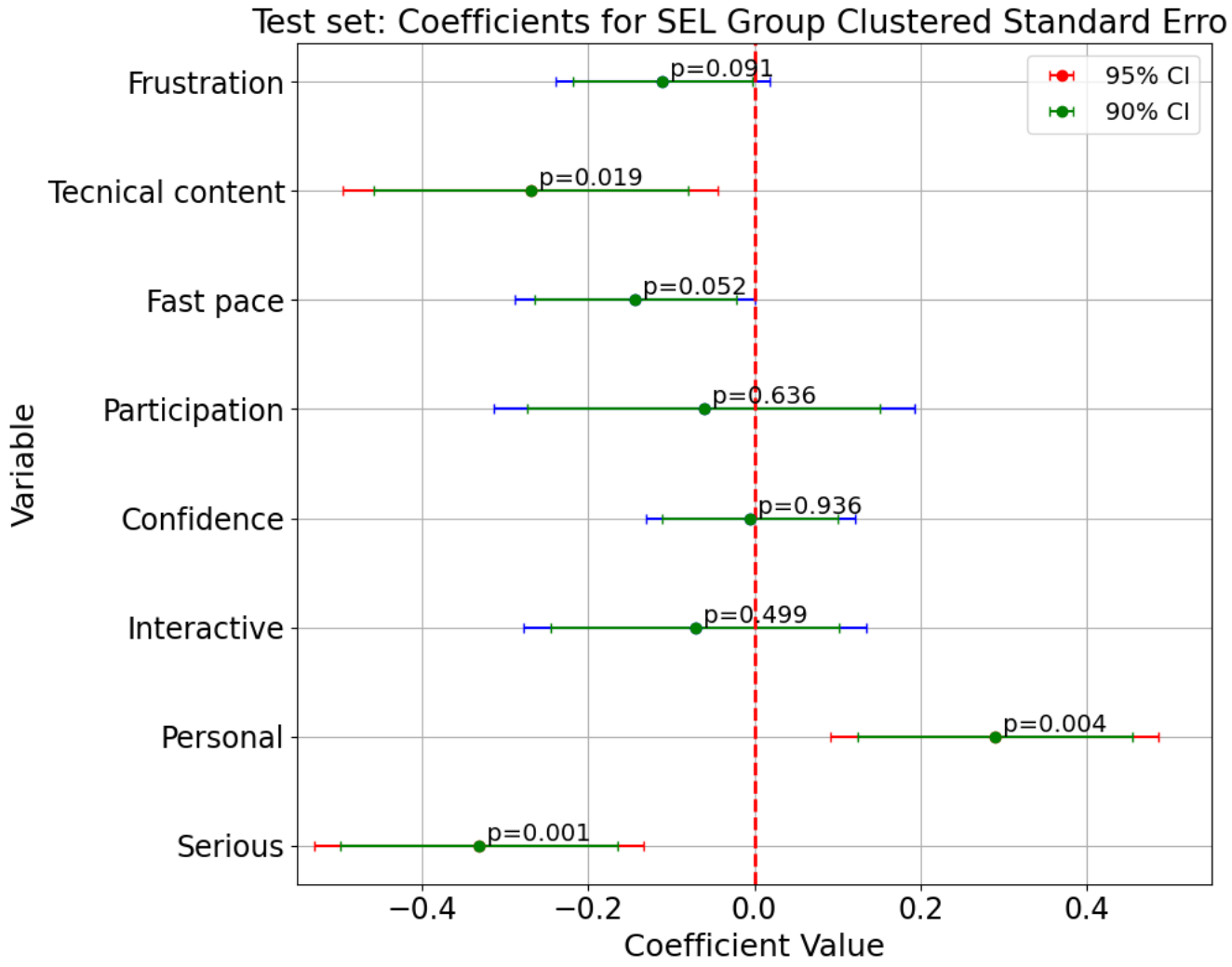
Appendix Table A.VII: Refined Prompt to Extract Differences between Types of Tutors

> You are a person walking on the street and someone asks you to compare two types of tutoring interactions.
>
> Your task is to identify how class A differs from class B based on the provided text A and B are random tags.
>
> IMPORTANT INSTRUCTIONS: - For each domain, provide exactly 3 qualitative differences (1-3 words each) - Each difference should clearly show HOW classes differ - Focus on observable behaviors within each specific domain
>
> DOMAINS TO COMPARE: 1. Tutor-Student Interactions: How the tutor and student interact with each other 2. Student Participation: How students engage in the class 3. Discussion Content: The topics of what is discussed 4. Discussion Mood: The emotional atmosphere of the session 5. Other differences that you observe
>
> - If you can't find 3 clear differences in a domain, use your best judgment or note "no clear difference" - Be specific and directional in your comparisons
>
> Please analyze the following text and return your response as a valid JSON object that matches this schema:

Notes: This table shows a refined prompt used with Claude 3.5 to extract differences between pairwise comparisons of *morphed* phrases.

A.8 Labeled figure with clustered standard errors test set

Appendix Figure A.II: Extracted Features



Appendix Figure A.III: Test set

Notes: In this figure we show the coefficients resulting from Equation 5, where s_r is the standardized score for each extracted label. For the test sample. We plot standardized treatment effects; the standard errors in this regression are clustered at the tutor level, the level of randomization and the bars represent 95% and 90% confidence intervals.

A.9 Sparse autoencoders for hypothesis generation

As a robustness check and complementary validation, we apply the sparse autoencoder hypothesis generation method of Movva et al. (2025) to the same classroom transcripts and compare the resulting hypotheses to those generated by our morphing procedure. The two methods differ in their approach: while our procedure generates controlled stylistic variations of extreme snippets and uses an LLM to name recurring differences, Movva et al. (2025) decompose text embeddings into sparse, interpretable features using a sparse autoencoder and identify which features differ most systematically between groups. Despite these methodological differences, the hypotheses generated by the two approaches are broadly consistent,

providing mutual validation.

Table A.9 reports the hypothesis-level results from the Movva et al. (2025) method on the held-out test set (AUROC = 0.581). The two hypotheses that are statistically significant at the 0.5% level are directly interpretable and align closely with our main findings. First, sessions that begin with greetings and well-being inquiries are more prevalent in the SEL condition (regression coefficient 0.062, $p = 0.005$, prevalence 24.5%), consistent with our finding of more *personal* topics and the warmer rapport documented in Section 4.2. Second, discussions of school vacation plans or experiences are more common in SEL classes (coefficient 0.088, $p = 0.078$, prevalence 3.3%), again reflecting the more personal and less formal tone identified by our procedure. On the other side, snippets that explain the *tarea de honestidad* activity — the name with which we refer to the endline evaluation — are strongly associated with the control group (coefficient -0.167, $p < 0.001$).

Several other patterns in the table are consistent with our findings without reaching conventional significance thresholds. Snippets involving step-by-step mathematical operations such as fractions (coefficient -0.056, $p = 0.011$) and explanations of algebraic variables (coefficient -0.064, $p = 0.015$) are less prevalent in the SEL condition, consistent with our finding of less *technical content* in SEL classes.

Movva et al. (2025)

Hypothesis	Sep. Score	Sep. p-val	Reg. Coef	Reg. p-val	Prev.
Mentions the names of students	0.168	0.101	0.123	0.229	0.007
Discusses plans for or experiences during school vacations	0.130	0.008	0.088	0.078	0.033
Mentions sending a link to students and waiting for them to join	0.120	0.012	0.171	0.001	0.034
Session begins with greetings and well-being inquiries	0.094	0.000	0.062	0.005	0.245
Asks students what topics they are studying in school	0.057	0.021	-0.001	0.979	0.149
Tutor asks explicitly what they are seeing in math at school	0.051	0.025	0.035	0.220	0.179
Mentions technical difficulties (audio/connection issues)	0.038	0.063	0.019	0.360	0.234
Instructs students to gather materials for activities	0.024	0.756	0.052	0.501	0.013
Practices or recites multiplication tables (especially 8)	0.000	1.000	0.016	0.738	0.035
Mentions a student completing exercises	-0.009	0.892	0.002	0.979	0.017
Discusses recording or sharing session recordings	-0.053	0.305	-0.069	0.185	0.030
Explains formulas for probability or proportionality	-0.056	0.035	-0.017	0.574	0.126
Explains step-by-step math operations (fractions, etc.)	-0.056	0.011	-0.026	0.313	0.199
Asks if exercises helped students	-0.063	0.143	-0.090	0.039	0.043
Explains use of variables in algebra	-0.064	0.015	-0.044	0.104	0.126
Mentions how to write or redact something	-0.065	0.008	-0.030	0.238	0.145
Explains 'tarea de honestidad' or related activity	-0.173	0.000	-0.167	0.000	0.057
Guides login into an educational portal	-0.189	0.007	-0.120	0.151	0.016

Appendix Table A.VIII: **Hypothesis-level results on holdout set (AUROC = 0.581, AUPRC = 0.569).** Hypotheses are automatically discovered behavioral patterns extracted from tutoring session transcripts using HypotheSAEs (Movva et al., 2025), which trains a sparse autoencoder on transcript embeddings and uses an LLM to translate learned features into natural language descriptions. *Separation Score* measures the difference in hypothesis activation between positive and negative outcome sessions (positive = more prevalent in SEL sessions); *Sep. p-value* is from permutation testing. *Regression Coefficient* reflects each hypothesis's association with SEL classes in a multivariate logistic regression controlling for all other hypotheses; *Prevalence* is the share of holdout sessions where each behavior is detected. Hypotheses are sorted by separation score: relational and rapport-building behaviors (e.g., greeting students, using names, connecting to school curriculum) are positively associated with SEL classes, while formal and abstract instructional patterns (e.g., step-by-step algebra, portal-based activities) show negative associations.