Teaching Practices, Behavioral Change, and Socio-Emotional Skill Formation *

CATERINA CALSAMIGLIA
Barcelona Supercomputing Center

GIACOMO DE GIORGI
University of Geneva, IEE/GSEM

LAIA NAVARRO-SOLA
IIES, Stockholm University

ANDREA SALVATI
Aarhus University

ECE YAGMAN
Universitat Autonoma de Barcelona

March 23, 2026

PRELIMINARY AND INCOMPLETE
Please do not distribute!

Abstract

Schools shape skills that matter beyond academic achievement. Teachers are central to this process, yet whether these effects arise from fixed teacher attributes or from malleable instructional practices remains unclear. We investigate this question through a randomized intervention in Spanish secondary schools. The program equips treated teachers with tools to observe students' socio-emotional skills and provide structured feedback during routine instruction, thus generating exogenous variation in instructional practices. We combine external observations of student behavior with independent measures of teaching practices, survey measures of socio-emotional skills, and administrative records of grades and absences. The intervention improves observed behaviors among disadvantaged students and those with weaker baseline skills. Treated teachers shift instruction toward more interactive forms and reduce behavioral correction. For disadvantaged students, these behavioral gains reduce absenteeism but do not raise short-run grades, while more advantaged students experience modest grade improvements despite little observable behavioral change. These patterns reveal that instructional practices causally shape socio-emotional skill formation, with behavioral and academic margins responding asymmetrically across student populations. Average treatment effects substantially understate welfare-relevant improvements and obscure how the same instructional change operates through distinct mechanisms depending on students' baseline conditions.

*Corresponding author: laia.navarrosola@iies.su.se. We gratefully acknowledge financial support from the Swiss National Science Foundation (grant no. 100018_207707). One author is a co-founder of the methodology studied and recused herself from data analysis and paper writing; a formal statement is included in the appendix.

I Introduction

Schools shape skills that matter well beyond academic achievement. Through daily classroom interactions, students develop socio-emotional skills—such as cooperation, responsibility, and perseverance—that contribute to educational attainment and generate persistent effects on labor-market and life outcomes (e.g., Almlund et al., 2011; Sorrenti et al., 2024). Teachers are central to this process: their instructional practices, feedback, and expectations shape the behaviors that express and reinforce these skills and, in turn, directly influence academic performance (Jackson, 2018; Kraft, 2019).

While a growing body of research demonstrates that teachers significantly affect student behaviors and socio-emotional skills , these observed effects do not reveal the mechanisms through which such changes occur. Teacher impacts may reflect persistent differences in teachers' interpersonal abilities, beliefs, or expectations, or they may operate through instructional and pedagogical practices that shape how attention is allocated, lessons are structured, and feedback is delivered. Among these channels, instructional practice is particularly amenable to policy intervention: unlike fixed teacher attributes, it can be modified within existing classrooms and is often the target of professional development efforts. Yet instructional choices are endogenous to teacher characteristics and classroom context, often correlating with broader dimensions of teacher quality (see e.g., Aucejo et al., 2022; Todd and Wolpin, 2018). Establishing whether and how instructional practices causally shape socio-emotional skills is therefore central to policy-design—informing, for instance, whether interventions should prioritize recruiting teachers with particular attributes or instead focus on changing instructional practices within existing classrooms.

Understanding these mechanisms, however, requires directly observing student behaviors through which socio-emotional skills are expressed in classroom settings. Most existing evidence instead relies on academic outcomes such as grades, attendance, or disciplinary incidents, or on narrow skill measures tied to standalone interventions. These indicators capture classroom behaviors only indirectly, making it difficult to identify whether changes in teaching affect specific student actions through which these skills develop. A comprehensive account of socio-emotional skill formation therefore requires observing student behaviors directly as they unfold in the classroom.

In this paper, we investigate whether and how instructional practices affect socio-emotional

skill formation inside classrooms and how these translate into academic outcomes. We address these questions by randomizing access to a teacher training and mentoring program implemented through a clustered randomized controlled trial in Spanish secondary schools. The program generates experimental variation in expected instructional practices by equipping treated teachers with tools to systematically observe students' socio-emotional skills and provide structured, behavior-focused feedback during routine lessons. It introduces no new academic content, requires no additional instructional time, and leaves formal assessment unchanged, thereby generating exogenous variation in how instruction is delivered within existing classroom routines.

To trace the impacts of this intervention, we combine data on external observations of student behavior—measured both during regular instruction and in a standardized group task—with independent observations and assessments of teaching practices, survey measures of students' socio-emotional skills, and administrative records on attendance and grades. This design allows us to examine whether changes in instructional practice alter observable classroom behavior and how those behavioral adjustments map into downstream academic outcomes.

Average effects on socio-emotional skills and academic outcomes are close to zero when pooling students across classrooms. However, impacts on observed behavior are strongly heterogeneous: Improvements are concentrated among students from more disadvantaged socio-economic backgrounds and those with weaker baseline socio-emotional skills, declining monotonically with socio-economic advantage and being close to zero for more advantaged students. For disadvantaged students, the intervention generates sizable and robust gains in observable behaviors related to autonomy and responsibility, detected in both the standardized activity and regular classroom observations. In contrast, self-reported measures of socio-emotional skills show limited movement on average.

The intervention generates clear and structured changes in instructional practice, with effects that are context dependent. Observations from external observers show that treated teachers increase interactive forms of instruction and reduce time spent correcting student behavior, without reducing time devoted to academic content on average. In classrooms serving more disadvantaged students, teachers primarily shift the mode of instruction, moving away from content- and output-focused interactions toward more interactive and dialogic engagement. In more advantaged classrooms, changes are concentrated in regulation of learning and behaviors, with greater encour-

agement of student agency and motivational support and less reliance on direct behavioral correction. Students' perceptions of teaching practices align with these patterns, with improvements in teacher effort concentrated among disadvantaged students. Teachers' self-reported time use shows no systematic changes across most instructional activities, with only weak evidence of reduced time spent addressing disciplinary issues.

Downstream educational outcomes respond asymmetrically. Among disadvantaged students, improved classroom behavior is associated with meaningful reductions in absenteeism but does not translate into higher grades in the short run. Among more advantaged students, academic performance improves despite little change in observed behavior, while attendance weakly declines. These patterns indicate that changes in instructional practice do not map uniformly into standard academic outcomes, and that different outcome margins adjust for different students depending on baseline constraints and institutional features.

Taken together, the results show that instructional practices can change the behaviors related to socio-emotional skills inside the classroom, but these changes translate imperfectly into conventional academic outcomes. The same instructional shift yields sizable behavioral and attendance gains for disadvantaged students while generating modest grade improvements for more advantaged students with little observable behavioral change. This heterogeneity implies that average treatment effects on grades or test scores can understate welfare-relevant improvements and obscure the margins through which classrooms produce human capital.

This paper first contributes to the human capital literature by providing direct evidence on how socio-emotional skills are produced inside classrooms. A large body of work establishes that socio-emotional skills are important determinants of educational and labor-market outcomes and are malleable during childhood and adolescence (Heckman et al., 2006; Cunha and Heckman, 2007; Cunha et al., 2010; Heckman and Kautz, 2012; Kautz et al., 2014; Alan et al., 2019; Sorrenti et al., 2024). However, most existing evidence infers skill formation from self-reports, incentivized tasks, or downstream academic outcomes, rather than observing how student behavior changes during routine instruction. This creates interpretation challenges because commonly used "non-cognitive" measures capture different constructs and contexts and often rely on latent skill indices (Borghans et al., 2008; Almlund et al., 2011; Humphries and Kosse, 2017; Attanasio et al., 2020; Jackson, 2018). By combining experimental variation in teaching practices with external, behavior-based

observation of students, we directly observe the behavioral margins through which socio-emotional skills adjust during instruction. We document that behavioral changes emerge before—and do not mechanically map into—changes in grades or attendance. This evidence could help explain why socio-emotional interventions can generate heterogeneous or muted academic effects even when meaningful behavioral change occurs inside the classroom, helping reconcile mixed findings in the literature (Heckman et al., 2010, 2013; Chetty et al., 2014; Sorrenti et al., 2024).[1]

Second, the paper contributes to the literature on teacher-implemented interventions by showing that socio-emotional skill formation can be shifted through instructional practices, even when curriculum content, instructional time, class size, and formal assessment remain fixed. Most socio-emotional learning interventions studied in economics operate as standalone curricula or add-on activities delivered alongside regular instruction (Alan et al., 2019; Santos et al., 2024; Alan and Mumcu, 2022; Sorrenti et al., 2024; Huillery et al., 2025; Gupta, 2025). In contrast, the intervention we study embeds socio-emotional learning directly into routine classroom practice by training teachers to systematically observe student behavior and provide feedback during regular lessons. This design isolates reallocations in how instruction is delivered from expansions in formal instructional inputs, allowing us to generate exogenous variation in instructional practice within existing classroom routines.

While a large literature establishes that teachers matter for student outcomes (Hanushek, 2003; Rockoff, 2004; Dobbie and Fryer Jr., 2013; Kraft et al., 2018), the specific practices through which these effects operate are often inferred rather than directly observed. At the same time, evidence on teacher professional development highlights the difficulty of changing classroom practice and the importance of program design, ongoing support, and teacher engagement (Garet et al., 2001; Popova et al., 2022; Cilliers et al., 2022; Beg et al., 2022; Oreopoulos et al., 2024). By combining experimental variation with direct observation of instructional practices and student behavior, this paper provides causal evidence on whether and how reshaping routine instruction can meaningfully

[1] A related contribution concerns measurement. A well-established literature emphasizes that socio-emotional skills are latent constructs and that commonly used measures—such as self-reports or teacher assessments—often differ in what they capture and in their predictive content (Borghans et al., 2008; Almlund et al., 2011; Humphries and Kosse, 2017; Malamud et al., 2024). To address this concern, we rely on trained external observers and behavior-based measures collected in both standardized and naturalistic classroom settings. This approach complements survey-based and administrative data and allows us to observe skill-relevant behavior as it unfolds during instruction, providing a triangulation of outcomes that strengthens interpretation and credibility without relying on any single measurement modality.

alter socio-emotional skill formation.

Third, the paper contributes to the literature on socio-emotional skill formation by documenting systematic heterogeneity in behavioral and academic responses to instructional change, and by linking this heterogeneity to observable classroom mechanisms. A growing body of work shows that socio-emotional skill interventions often produce uneven effects across students, with average treatment effects masking important distributional patterns (Santos et al., 2024; Sorrenti et al., 2024; Huillery et al., 2025). However, existing evidence typically focuses on final outcomes, leaving less understood why the same pedagogical change benefits some students but not others, or why different outcome dimensions respond asymmetrically.

We show that improvements in socio-emotional behavior are concentrated among students with weaker baseline skills and more disadvantaged backgrounds, consistent with models in which the productivity of instructional inputs depends on initial conditions and complementarities across skills (Cunha and Heckman, 2007; Todd and Wolpin, 2003). For these students, behavioral gains translate into meaningful reductions in absenteeism but not immediate grade improvements, whereas more advantaged students experience modest grade gains despite limited changes in observed behavior. This pattern is consistent with evidence that schooling features such as grading practices, tracking, and attendance rules, mediate how behavioral changes translate into academic outcomes (Lazear, 2001; Dobbie and Fryer Jr., 2013; Burgess et al., 2022; Malamud et al., 2024). By tracing adjustments in classroom behavior, attendance, and grades jointly, this paper shows that the same instructional change can generate distinct margins of response across students, highlighting the importance of behaviorally mediated channels for understanding inequality and policy targeting (Jackson, 2018; Chetty et al., 2011, 2014).

The rest of the paper proceeds as follows: Section II describes the intervention. Section III details the experimental design and implementation, and Section IV presents the data sources and measurement strategy. Section V describes the sample and assesses baseline balance. Section VI outlines the identification strategy and estimation approach. Section VII presents the effects on socio-emotional skill–related behaviors, Section VIII examines impacts on teacher practices and classroom processes, and Section IX analyzes downstream academic outcomes. Section X concludes.

II *Pentabilities*: A Pedagogical Program for Socio-Emotional Learning

The intervention evaluated in this study, *Pentabilities*, is a pedagogical program that trains and mentors teachers to integrate socio-emotional skill formation into standard classroom instruction. The program operationalizes socio-emotional learning (SEL) through a structured pedagogy.[2]

A central feature of the intervention is the regular collection of behavior-based measures of students' socio-emotional skills by the teacher, which teachers use to generate feedback during the learning process. Importantly, the intervention does not alter instructional time, curriculum content, class size, or subject coverage. Instead, it changes how teachers observe, assess, and provide feedback on students' socio-emotional skills within existing classroom activities.

Pentabilities is an integrated package combining teacher training, mentoring, structured measurement of student behavior, and feedback supported by digital tools; these components are not independently randomized and are designed to be mutually reinforcing.[3] The program is delivered to schools as a professional development program for teachers.

The intervention is designed to affect teacher practices and student behavior within existing instruction. Section II.1 describes the pedagogical foundations of the approach, while Section II.2 outlines a conceptual framework that links these instructional changes to observed behavior and educational outcomes. Sections II.3 and II.4 describe, respectively, the day-to-day operationalization of the program into classrooms, and the teacher professional development package.

II.1 The Socio-Emotional Learning Pedagogy

The program's pedagogical approach is organized around two core principles: (i) the use of observable behaviors as proxies for latent socio-emotional skills, and (ii) the use of repeated, evidence-based feedback to influence skill development over time.

[2] "Structured pedagogy are those with a coherent package of investments that include lesson plans, learning materials, skills-based ongoing teacher training, and teacher mentoring that are carefully coordinated to reinforce each other," and have been highlighted as highly cost-effective interventions (Akyeampong et al., 2023).

[3] Pentabilities was developed collaboratively by experienced educators and specialists, and it has been implemented over the past decade across multiple educational levels in Spain and other countries.

OBSERVABLE BEHAVIORS AS PROXIES FOR SKILLS. Socio-emotional skills are latent factors that are not directly observable, making them difficult to assess, track progress, and target through instruction. The first principle of the pedagogy addresses this challenge by focusing on observable classroom behaviors that reflect underlying socio-emotional skills, consistent with a well-established literature that treats behavior as an empirical manifestation of latent non-cognitive skills (Heckman and Kautz, 2012; Jackson, 2018).

The pedagogy targets 35 specific, observable classroom behaviors, grouped into five broad socio-emotional skill domains: responsibility, autonomy and initiative, cooperation, thinking abilities, and emotional management.[4][5] These behaviors translate abstract socio-emotional constructs into concrete, measurable outcomes that can be systematically observed and discussed in classroom settings.

EVIDENCE-BASED FEEDBACK. The second principle is the use of repeated evidence-based feedback as the primary mechanism for skill development. Teachers regularly collect evidence on targeted behaviors and use this information to provide feedback to students during the learning process. Feedback is informational rather than evaluative, and is not tied to formal grades.

By making expectations explicit and providing repeated information on observed behavior, the pedagogy supports more precise communication between teachers and students about socio-emotional skills. Feedback is generated using standardized scales and is complemented by student self-assessments and peer evaluations, allowing multiple perspectives on behavior to be incorporated. This structure is intended to facilitate reflection and goal-setting within routine instruction.

II.2 Conceptual Framework: Instructional Inputs, Behaviors, and Outcomes

The pedagogical principles described above motivate a simple conceptual framework that links changes in instructional practices to student behavior and downstream educational outcomes. The framework is intended to discipline interpretation of the empirical results.

The framework views the classroom as an environment in which teachers allocate limited peda-

[4]These behaviors were identified using rubrics, personality questionnaires, and assessments commonly used in educational and professional settings (Soto et al., 2022).

[5]Examples include *"respects the rules"* (responsibility), *"incorporates what others say"* (cooperation), *"takes decisions to advance the project"* (autonomy and initiative), *"manages their emotions in conflict situations"* (emotional management), and *"plans and prioritizes tasks"* (thinking abilities). See Tables B2 and B3 for the full list of behaviors.

gogical attention across multiple instructionals tasks (such as academic instruction, socio-emotional development, and classroom management) and across students who differ in baseline skills, engagement, and external constraints. Within this setting, the intervention reshapes *how* instructional inputs are delivered during routine instruction, rather than *how much* instruction takes place.

INSTRUCTIONAL INPUTS AND THE MARGIN OF ADJUSTMENT. Teachers face binding time and attention constraints. Beyond delivering academic content, they must manage classroom interactions, provide feedback, and respond to behavioral or disciplinary issues. These inputs are not perfect substitutes: time devoted to maintaining order or resolving conflicts may crowd out individualized attention, while relational and feedback-oriented practices may increase engagement at relatively low marginal cost.

The intervention studied in this paper does not increase instructional time, introduce new academic materials, or alter curricular requirements. Instead, by embedding systematic behavioral observation and feedback into standard classroom activities, it changes the composition and effectiveness of pedagogical inputs. Conceptually, the intervention raises the salience and productivity of relational and feedback-based practices, and may reduce the need for disciplinary intervention, without requiring an increase in total teacher effort.

STUDENT BEHAVIOR AS AN INTERMEDIATE MARGIN. Instructional inputs affect students primarily through observable classroom behaviors. Forms of classroom engagement, participation, cooperation, and task-oriented behavior represent intermediate inputs into both learning and participation and are directly shaped by teacher–student interactions and peer dynamics. These behaviors are observable to external observers in both structured activities and regular classroom settings.

A key distinction in the framework is between changes in observable behavior and changes in latent traits. Broad personality traits may evolve slowly or be imperfectly captured by surveys, whereas observable behaviors can respond more immediately to changes in instructional practices. As a result, interventions that modify classroom interactions may generate detectable changes in behavior even in the absence of shifts in global self-reported measures. This distinction motivates the use of behavior-based measures as primary outcomes in the empirical analysis, with survey-based measures playing a complementary role.

A natural mechanism through which such behavioral changes may arise is repeated feedback based on observable actions. By making expectations explicit and providing information during routine instruction, feedback can shape attention, effort, and classroom interactions through a range of informational and relational channels.

HETEROGENEOUS COMPLEMENTARITIES IN SKILL FORMATION. An important implication of this framework is heterogeneity in the returns to instructional inputs. Students enter the classroom with different baseline levels of engagement, regulation, and external support. For students who face constraints in participation or classroom integration (conditions that are more prevalent among students from disadvantaged socio-economic backgrounds) the marginal productivity of relational and feedback-oriented pedagogy may be higher.

In contrast, students with stronger baseline skills may exhibit smaller behavioral responses to the same instructional changes, even though they may be better positioned to translate engagement into academic performance. These heterogeneous complementarities imply that average treatment effects may obscure offsetting gains across students. The framework therefore highlights conditions under which improvements in observable socio-emotional behavior may be concentrated among students starting from weaker initial conditions, even when average classroom-level effects are modest.

FROM BEHAVIOR TO EDUCATIONAL OUTCOMES. Observable behaviors serve as inputs into downstream educational outcomes such as attendance and academic performance. However, the mapping from behavior to outcomes might be neither linear nor uniform across students. Improvements in engagement and responsibility may reduce absenteeism among students who were previously disengaged, while improvements in academic performance may occur primarily among students who are already close to the margin at which performance is rewarded through formal assessment. Institutional features (such as grading practices, subject-specific evaluation, and attendance requirements) further mediate this mapping. As a result, changes in behavior need not generate parallel movements in grades and attendance, and different outcomes may respond for different students.

Taken together, this framework highlights a set of empirical patterns that guide the analysis

in the following sections. It suggests that instructional interventions of this type may first affect observable classroom behavior, that such effects may be heterogeneous across students depending on baseline constraints, and that downstream educational outcomes may respond asymmetrically across dimensions and groups. Sections VII through IX evaluate these implications by tracing the effects of the intervention from instructional inputs, to observed behavior and downstream educational outcomes.

II.3 Operationalization in the Classroom

Teachers implement the pedagogy through structured instructional cycles embedded within their standard curriculum. Each cycle consists of three stages: planning, observation and data collection, and feedback. While teachers retain discretion in the classroom activity design, the targeted behaviors, measurement scales, and feedback structure are standardized across classrooms.

During the planning stage, teachers select socio-emotional skills to target and design classroom activities intended to elicit the corresponding behaviors. For example, to develop the behavior "listens to others" within the cooperation domain, a teacher might design small-group discussions that require turn-taking and active listening. During instructional time, teachers observe and record instances of targeted behaviors using digital tools. Observations are complemented by student self-assessments and peer evaluations. In the feedback stage, collected evidence is aggregated by the digital tools into individualized reports that teachers use to provide individual or group feedback and to set goals for subsequent cycles.[6]

Educational technology supports the implementation of these cycles by lowering the cost of repeated measurement and feedback. A mobile and web-based platform allows teachers, students, and peers to record observations of behaviors in real time, and automatically aggregates information into standardized feedback reports. By reducing administrative burden and standardizing data collection, the technology enables systematic monitoring of socio-emotional behavior within routine instruction.

This system operationalizes "formative assessment", defined as a teaching practice in which feedback is provided continuously during the learning process rather than solely at its conclusion (Sadler, 1989). While formative assessment is well established in education, it represents a mean-

[6]Figure A3 provides snapshots of the student report provided to teachers, and Figure A2 a snapshot of the educational technology.

ingful departure from the lecture-based teaching practices prevalent among teachers in the study, where feedback is infrequent and coarse.

II.4 The Teacher Professional Development Program

To ensure consistent implementation, Pentabilities includes a structured professional development component for teachers. The program consists of an initial training workshop, followed by ongoing individualized mentoring and two group sessions.

The program begins with an 8-hour training workshop, combining remote and in-person components, which introduces the pedagogy, the behavioral observations, and the use of the digital tools. Following the workshop, teachers receive monthly one-on-one mentoring from experienced educators trained in the Pentabilities methodology, supporting implementation of the instructional cycles. In addition, teachers participate in two group sessions with other teachers during the implementation period, which focus on interpreting behavioral data and sharing implementation experiences.

III Experimental Design and Implementation

We evaluate the intervention using a clustered randomized controlled trial conducted during the 2022–2023 academic year in 40 public and state-assisted secondary schools in Catalonia, Spain. Randomization occurred at the school-grade cluster level, with participating schools contributing at least two grades, one assigned to treatment and one to control. Treated grades received the teacher professional development program consisting of training, mentoring, and access to the digital tools to support socio-emotional skill formation, while control grades continued standard instructional practices during the study period. Baseline data were collected prior to randomization, and outcome data were collected at the end of the academic year, in May–June 2023. Our primary estimand is the intention-to-treat effect of assignment of school-grade clusters to the teacher professional development program.

SAMPLE SELECTION. Between June and September 2022, we recruited secondary schools in Catalonia through public outreach, including newsletters, municipal and local education networks, and referrals within local education communities. Participation was voluntary at both the school and teacher levels, and teachers opted into the study by applying to implement the intervention in their classrooms. Recruitment primarily targeted schools serving socio-economically disadvan-

taged student populations.

To guide recruitment, we relied on administrative and publicly available indicators of schools' socioeconomic composition such as the school disadvantage level, the school funding type (public, private, or state-assisted), and local-area level socio-economic indicators.[7]

School eligibility was also shaped by methodological considerations. To support within-school randomization, participating schools were required to contribute at least two grade levels to the study. Within each participating grade, teachers who volunteered to take part selected one or more classrooms in which they provided at least three hours of weekly instruction, either in a single subject or across multiple subjects, and in which they could feasibly implement the intervention. To avoid contamination across treatment arms, teachers were required to teach in only one participating grade. These classrooms defined the school-grade clusters used in the experimental design.

Since teacher participation was voluntary, all causal estimates identify the effect of assignment to the intervention among teachers who opted into participation. Because randomization occurred after teacher enrollment and baseline data collection, volunteer status does not threaten internal validity, though it may not generalize to populations with substantially different levels of baseline motivation.

INTERVENTION AND RANDOMIZATION. The intervention was evaluated using a clustered randomized controlled design. Randomization occurred at the school-grade cluster level within each participating school. Each school contributed at least two grade levels, with at least one grade assigned to treatment and one to control.

Random assignment was conducted after baseline data collection. We stratified randomization by school disadvantage level. Within each school, treatment was randomly assigned among the eligible grade clusters. This design ensures that treatment and control clusters are exposed to similar school-level disadvantage environments.

Volunteer teachers within the school-grade clusters assigned to the treatment group received the teacher professional development program, which included initial training, ongoing mentoring, and access to digital tools to support socio-emotional skill formation. Clusters assigned to the

[7] We used four categories classifying schools stemming from an administrative composite index used by the Catalan education authority as a proxy for school-level socioeconomic disadvantage. For more details on the index, see Grimaldo and López-Iñesta (2021).

control group continued standard instructional practices during the study period and did not receive any training, mentoring, or access to the intervention tools. Control teachers were offered the opportunity to participate in the training after the end of the study period.

Our primary estimand is the intention-to-treat effect of assignment to the teacher professional development program. The initial randomized sample consists of 88 school-grade clusters (4,763 students) in total. Due to budget constraints, outcome data was collected on a pre-specified subset of classroom groups within each school-grade cluster, selected prior to revealing treatment assignment. Selection rules were identical across treatment and control grades, ensuring that the measurement sample preserves the experimental balance induced by randomization. All analyses are conducted on this measurement sample of 2,068 students: 40 school-grade clusters (69 classrooms, 1,015 students) assigned to treatment and 40 school-grade clusters (72 classrooms, 1,053 students) assigned to control, reflecting the within-school randomization design. We discuss attrition in Section VI.

IMPLEMENTATION, COMPLIANCE, AND POTENTIAL SPILLOVERS. The intervention was implemented as planned, with teachers assigned to treated grades receiving an initial training workshop, individualized mentoring, and access to a digital platform supporting behavioral measurement and feedback.[8]

As is typical in professional development interventions, engagement intensity varied across treated teachers. While all treated teachers were offered identical access to training, mentoring, and digital tools, some made more intensive use of these components than others, and a small number of treated teachers exhibited minimal engagement. All primary analyses follow an intent-to-treat (ITT) framework and do not condition on realized take-up or usage. Control-grade teachers did not receive any training, mentoring, or access to the digital platform during the study period and interacted with the implementation team only for survey and observation logistics. Control teachers were offered access to the training after the conclusion of the study.

The experimental design limits scope for spillovers across treatment arms. Randomization occurred at the grade-within-school level, and participating teachers were required to teach in only

[8]Training and mentoring were delivered by experienced educators from *Empieza por Educar* (Teach for All), who were not involved in outcome measurement or data collection.

one study grade, preventing direct cross-grade contamination through shared instruction. Core components of the intervention (structured behavioral measurement, digital feedback tools, and mentoring) were accessible only to treated teachers. While informal interactions among teachers within the same school cannot be ruled out, any remaining spillovers would likely attenuate estimated treatment effects toward zero.

Accordingly, the estimated effects should be interpreted as intent-to-treat impacts of offering the teacher professional development program under realistic conditions of heterogeneous engagement.

STUDY TIMELINE. The study was conducted during the 2022-2023 academic year. Figure A1 summarizes the timeline of the intervention and data collection. Baseline data was collected between November 2022 and January 2023 and included student and teacher surveys as well as in-class observations. Randomization was conducted in late November 2022. Baseline survey for some schools continued through January; however, all baseline measures were collected prior to any disclosure of treatment status and before the start of teacher training or classroom implementation. Treatment status was disclosed to each school only after baseline data collection was finalized in that school.

Teachers assigned to treatment clusters began training in December 2022. Initial training workshops were completed by the end of January 2023, after which teachers started implementing the intervention in their classrooms. Program implementation took place from January-February to June 2023. Due to logistical constraints, the start of implementation was staggered across schools. Outcome data was collected at the end of the academic year, in May–June 2023, after all treatment clusters had been exposed to the intervention. Control clusters did not receive any training or program materials during the study period.

IV Data and Outcome Measures

We combine four complementary data sources to measure impacts on students and teachers, with primary outcomes based on externally observed behavior and administrative records, and secondary outcomes based on validated survey instruments. Data sources include: (i) student and teacher surveys administered in school by an external survey firm using tablets; (ii) repeated in-classroom observations conducted by trained external observers; (iii) behavior-based measures col-

lected during a standardized endline group activity; and (iv) administrative attendance and grade records provided by schools. For treated classrooms, we additionally collect implementation data from educational-technology logs and mentoring reports, which are used to characterize compliance and implementation intensity.

Classroom observations and the standardized activity were implemented by observers hired and managed by an independent organization (Empieza por Educar). Observers were blind to treatment assignment, did not participate in intervention delivery, and followed identical protocols in treatment and control schools. Baseline data was collected between November 2022 and January 2023, midline data between March and April 2023, and endline data between May and June 2023. Unless otherwise noted, outcomes are measured on the study's measurement sample.

Behavior-based outcomes are measured at baseline (classroom observations only) and endline (classroom observations and standardized activity). Survey-based outcomes are measured at baseline, midline, and endline, with some instruments administered at endline only to reduce respondent burden.

IV.1 Outcome Variables

We organize outcomes into student-centered educational outputs and teacher-centered educational inputs within a standard education production-function framework. Student outputs capture socio-emotional skills and related behaviors, and well-being and social relationships. Administrative grades and attendance are treated as downstream educational outputs that may respond to changes in socio-emotional skills and classroom practices. Teacher inputs capture instructional practices and time allocation, including practices that may mediate socio-emotional skill formation. Table B1 summarizes outcome families, data sources, timing, and sample coverage. Outcome measurement details are provided in Appendix B.

STUDENT-CENTERED OUTCOMES (EDUCATIONAL OUTPUTS). We measure socio-emotional skills using two complementary approaches. First, we construct behavior-based measures from external observations that capture the manifestation of targeted skills (autonomy and initiative, cooperation, emotional management, responsibility, and thinking abilities) through a structured framework of 35 observable behaviors. These behaviors are rated both during repeated classroom obser-

vations and during a standardized endline group activity administered outside the regular classroom setting. These measures constitute the primary outcomes of the analysis.

Second, we collect survey-based measures of socio-emotional skills using validated instruments and a 360-degree design combining student self-reports, randomly assigned peer assessments, and teacher assessments for randomly selected subsets of students. These survey measures are used to triangulate behavior-based outcomes and to study mechanisms and robustness.

We also measure student well-being and social relationships using additional survey modules, including student well-being and classroom friendship networks. Finally, we use administrative records on attendance and academic grades for the 2022–2023 academic year as downstream educational outcomes. Administrative coverage is incomplete: attendance records are available for 21/40 schools and term grade records for 31/40 schools. We document coverage in Section V.

TEACHER OUTCOMES (EDUCATIONAL INPUTS). We measure teacher practices and inputs using three sources: (i) student reports of instructional practices; (ii) repeated external observations of classroom practices using the Teacher Roles Observation Schedule (TROS) (Waxman et al., 1990); and (iii) teacher surveys measuring time use, motivation, teaching mindset, and socio-emotional skills. Each classroom was visited between four and six times over the academic year, symmetrically across treatment and control groups.

IMPLEMENTATION MEASURES (TREATMENT ONLY). For treated classrooms, we collect educational-technology usage logs and mentoring reports. These measures are used to characterize implementation intensity and compliance and are not analyzed as primary outcomes.

PRIMARY AND SECONDARY OUTCOMES. Following the pre-analysis plan, the primary outcomes of the study are behavior-based measures of student socio-emotional skills constructed from external observations, both during regular classroom instruction and during the standardized endline group activity. These measures are directly aligned with the skills targeted by the intervention and are least subject to self-report bias. Survey-based measures of student socio-emotional skills, well-being, and social relationships, as well as teacher-reported outcomes and administrative educational outcomes, are treated as secondary outcomes and are used to triangulate results, study

mechanisms, and assess robustness.

IV.2 Behavior-based Measures of Socio-Emotional Skills

A central contribution of the study is the use of externally observed, behavior-based measures aligned with the skills targeted by the intervention.

At endline, all students in the measurement sample participated in a standardized one-hour group activity conducted outside the regular classroom setting. The activity was designed to elicit socio-emotional behaviors in a common environment, reducing confounding from teacher-specific practices or classroom routines. External observers rated each student on a standardized five-point scale using a subset of 20 observable behaviors drawn from the broader behavioral framework. Students also completed brief self- and peer-assessments following the activity.

In parallel, external observers conducted repeated classroom observations between January and May 2023. Observations focused on randomly selected subsets of students within each classroom and captured the manifestation of 35 targeted behaviors during regular instruction. Observations were conducted symmetrically in treatment and control groups.

The use of a structured task complements in-classroom observations and aligns with recent experimental studies evaluating socio-emotional skills in school settings (e.g., Alan et al. (2019), Alan and Mumcu (2022) and Alan and Kubilay (2025)). As in this literature, the activity is designed to elicit meaningful behavioral variation while maintaining strict standardization and observer independence. Details on the behavioral framework and coding rules, standardized activity and in-class observation protocols are provided in Appendices B.2, B.3 and B.4, respectively.

IV.3 Survey Instruments

Student and teacher surveys provide complementary measures of socio-emotional skills, perceptions, and instructional practices. These measures are not designed to replace behavior-based outcomes, but to triangulate them using validated self-, peer-, and teacher-reported instruments and to support mechanism analysis.

Student surveys measure socio-emotional skills, well-being, and social relationships using established instruments administered at baseline, midline, and endline. Peer assessments are based on random assignment of classmates, and teacher assessments are collected for randomly selected sub-

sets of students, limiting scope for self-selection. Teacher surveys capture instructional practices, time use, motivation, teaching mindset, and socio-emotional skills, and are used to characterize changes in educational inputs and potential mechanisms.

All surveys were administered in either Catalan or Spanish, according to each respondent's preferred language. A complete description of survey instruments, versions, and timing is provided in Appendix B.5.

IV.4 Construction of Outcome Indices

To summarize information across multiple related outcomes and to address concerns about multiple hypothesis testing, we construct standardized outcome indices for each pre-specified domain. Within each domain, individual outcome measures are first oriented so that higher values correspond to better outcomes. Each component is then standardized using the mean and standard deviation of the control group, at baseline when a baseline measure is available and at endline otherwise. Domain-specific indices are constructed as the unweighted mean of the standardized components within each domain, requiring at least one non-missing component. The resulting indices are re-standardized to have mean zero and unit variance in the control group.

We construct separate indices for student socio-emotional skills measured through behavior-based observations. Treatment effects are estimated on each of the standardized socio-emotional skill domain indices, with individual behaviors analyzed separately in supplementary analyses.

For behavior-based outcomes, missing values arise only when a behavior could not be meaningfully observed, as defined in the coding protocol. For survey-based outcomes, item non-response is limited and not systematically related to treatment assignment. Indices are constructed using all available components, requiring at least one non-missing item per domain.

IV.5 Main Heterogeneity: Socio-Economic Background

The intervention was designed and evaluated with a particular focus on students' socio-economic background. From a conceptual perspective, socio-emotional skill formation may interact with students' home environments and prior opportunities, potentially generating heterogeneous impacts across socio-economic groups due to complementarities with family and home inputs. From a policy perspective, the program was implemented in schools serving socio-economically disadvan-

taged populations, reflecting an explicit objective of the funder (the Spanish Ministry of Social Inclusion) of improving equity and inclusion in educational outcomes. As defined above, school socio-economic characteristics were also used as a criteria for study participation at the school level, and we stratified randomization at this level.

Accordingly, heterogeneity by socio-economic background was pre-specified in the analysis. In Sections VII through IX, we examine differential treatment effects along a pre-specified household-level socio-economic index constructed from student survey data, capturing within-school variation.[9]

V Sample Characteristics and Baseline Balance

This section describes the experimental sample and assesses balance between treatment and control groups prior to the intervention. Treatment was randomized at the grade-within-school level, while outcomes are measured at the student and teacher levels. All descriptive statistics and balance assessments respect this experimental design.

SAMPLE CHARACTERISTICS. The study includes 39 public and state-assisted secondary schools in Catalonia, contributing a total of 80 grades and 148 classrooms. On average, each participating grade includes 2.7 classrooms (s.d. = 0.9) and 2.3 participating teachers.

The student analysis sample consists of 2,068 students with parental consent and non-missing outcome data. Students are between 11 and 17 years old, with the majority concentrated between ages 12 and 14. Approximately 87 percent of students were born in Spain, and 61 percent report that both caregivers were born in Spain.

The teacher sample consists of 116 teachers who completed baseline surveys. Teachers teach a wide range of subjects and vary substantially in experience. Approximately 68 percent of teachers are female, and most teachers hold either a bachelor's degree (43 percent) or a master's degree (46 percent). Teaching experience ranges widely, from early-career to more than 15 years.

Descriptive statistics for students and teachers are reported in Appendix C.

[9] Two complementary dimensions were also pre-specified: (i) school-level disadvantage indicators derived from administrative sources, capturing differences in school environments; and (ii) caregiver immigration background, capturing dimensions of social inclusion. These three measures were selected prior to analysis and reflect both the design of the intervention and its intended policy relevance.

Note that different outcome families rely on partially overlapping samples due to survey participation, parental consent requirements, and administrative data availability. For example, administrative records on attendance and grades are available for a subset of schools; coverage varies by outcome and is documented in Appendix C. Section VI describes how consent and attrition are handled in the estimation sample and assesses potential threats to identification.

BASELINE BALANCE. We assess baseline balance by comparing treatment and control groups along pre-treatment student, teacher, and school–grade characteristics. Balance is assessed by regressing pre-treatment covariates on treatment assignment, including stratification fixed effects and clustering standard errors at the school–grade level. Joint balance tests are also reported. Student-level and teacher-level balance are exclusively assessed on the measurement sample used in the analysis, due to data availability restrictions. Detailed balance tables and test statistics are provided in Appendix C.

Across student, teacher, and school–grade characteristics, we find almost no systematic or economically meaningful differences between treatment and control groups at baseline.

A small number of baseline variables exhibit minor imbalances; following the pre-analysis plan, these variables are included as controls in all main outcome regressions.

VI Identification and Estimation

ECONOMETRIC SPECIFICATION. Our estimand is the intent-to-treat (ITT) effect of assignment of a grade-within-school cluster to the teacher professional development program on student- and teacher-level outcomes, regardless of realized take-up or implementation intensity.

The main empirical specification follows an ANCOVA framework and is fully saturated with respect to the experimental design and pre-specified covariates. For individual i in grade g at school s, we estimate:

$$y_{igs} = \alpha + \beta T_{gs} + \gamma y_{igs0} + \delta_s + \eta_g + \sum_{w=1}^{W} \lambda^w + \sum_{k=1}^{K} \nu_k x^k_{igs0} + \varepsilon_{igs} \tag{1}$$

where y_{igs} denotes the outcome of interest for individual i (student or teacher) in grade g at school s, T_{gs} is an indicator equal to one if grade g in school s was assigned to treatment, and

y_{igs0} is the baseline value of the outcome when available. The coefficient β captures the intent-to-treat effect of assignment to the intervention. For outcomes measured only at endline, the baseline outcome y_{igs0} is omitted.[10]

The specification includes school fixed effects (δ_s), fixed effects for the randomization strata defined by school disadvantage level (η_g), and week-of-measurement fixed effects for outcomes collected over extended fieldwork periods (λ^w). We also include a pre-specified set of individual-level socio-demographic covariates and a small number of baseline variables that exhibited imbalance across treatment arms. All external observer-based outcomes include observer fixed effects to account for systematic differences in rating styles across observers.[11]

We cluster standard errors at the school–grade level, the unit of randomization (Abadie et al., 2023). Teacher outcomes derived from teacher surveys are analyzed at the teacher level, while outcomes based on student perceptions of teacher practices are aggregated to the teacher level; in all cases, inference is based on clustering at the school–grade level.

Given the large number of related outcomes, we address multiple hypothesis testing by reporting sharpened false discovery rate (FDR) q-values within pre-specified outcome families (Anderson, 2008). Outcome families were defined ex-ante and correspond to theoretically coherent domains rather than statistical groupings. Both raw p-values and FDR-adjusted q-values are reported in the tables.

The results presented in the paper are based on this baseline specification. Appendix D reports alternative specifications and robustness checks. Estimated treatment effects are stable across these exercises.

VII Impacts on Socio-Emotional Skill Formation

VII.1 Behavior-Based Socio-Emotional Skill Measures

We measure socio-emotional skill–relevant behavior using trained external observers in two settings: a standardized group activity administered at endline and in-classroom observations conducted during regular instruction. These behavior-based measures constitute the primary outcomes

[10] Baseline outcomes are available for classroom-observed behaviors and survey-based socio-emotional measures, but not for standardized group activity outcomes or administrative outcomes collected.

[11] All fixed effects and controls were pre-specified in the pre-analysis plan. See Appendix D.1 for details.

of the analysis because they capture observable manifestations of socio-emotional skills as they are expressed in action and do not rely on self-reports. Outcomes are constructed as domain indices based on observed behaviors and standardized using the control group mean and standard deviation. All estimates are reported on an intent-to-treat basis.

Table 1: Impacts on Behavior-Based Socio-Emotional Skills Assessed by External Observers

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Autonomy		**Cooperation**		**Emotional Management**		**Responsibility**		**Thinking Abilities**	
	Activity	Class	Activity	Class	Activity	Class	Activity	Class	Activity	Class
Treated	0.039	0.043	0.068	0.042	0.084*	0.158*	0.056	0.086	-0.044	0.188*
	(0.034)	(0.077)	(0.048)	(0.093)	(0.048)	(0.093)	(0.051)	(0.071)	(0.047)	(0.098)
Control mean	0.03	0.02	0.02	0.01	0.01	-0.00	0.03	0.01	0.05	0.03
Observations	1641	822	1644	736	1631	579	1632	965	1633	524
MHT q-val.	0.47	0.63	0.47	0.63	0.47	0.47	0.47	0.47	0.47	0.47

Notes: This table reports intent-to-treat estimates of the effect of assignment to the intervention on behavior-based socio-emotional skill indices measured by external observers. Outcomes are standardized using the control group mean and standard deviation. Each column corresponds to a distinct skill domain measured either during a standardized group activity administered at endline or through in-classroom observations conducted during regular instruction. All specifications include school fixed effects and fixed effects for the randomization strata defined by school disadvantage level. Week-of-measurement fixed effects are included. Regressions control for pre-specified student socio-demographic characteristics and baseline covariates that exhibited imbalance across treatment arms. Baseline measures of the outcome are not available for these indices. Robust standard errors are reported in parentheses and are clustered at the school-grade level. The row labeled "MHT q-val." reports sharpened false discovery rate q-values accounting for multiple hypothesis testing within each outcome domain. Statistical significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

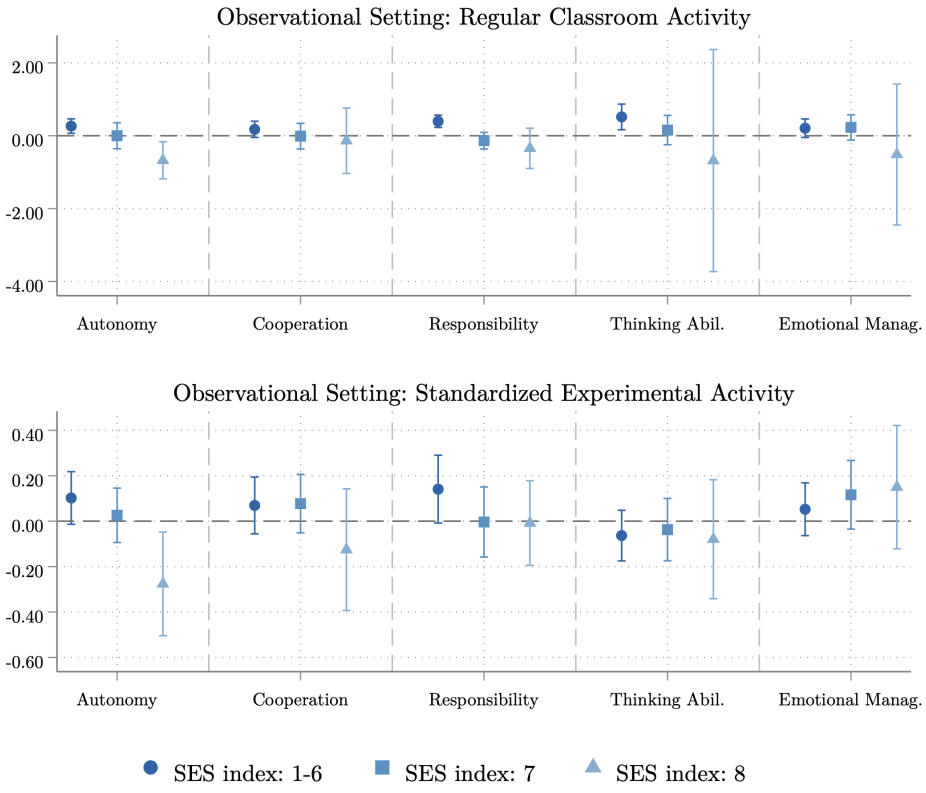
AVERAGE EFFECTS. Table 1 reports treatment effects on five socio-emotional domains —autonomy, cooperation, emotional management, responsibility, and thinking abilities— measured in both observational settings. Across domains, average effects are small and statistically indistinguishable from zero, and none remain significant after adjusting for multiple hypothesis testing. This pattern is similar across the standardized activity and regular classroom observations, indicating no detectable average change in observed behavior when pooling students across different charcteristics

.

DISTRIBUTIONAL IMPACTS BY STUDENT CHARACTERISTICS. Average treatment effects mask substantial heterogeneity in responses across students. Figure 1 plots treatment effects by students' pre-treatment socio-economic background.[12] The figure reveals a clear gradient: treatment effects are largest among students from more disadvantaged backgrounds and attenuate somewhat monotonically with socio-economic advantage. This pattern is present in both observational settings

[12]Section IV.5 provides details on the construction of the socio-economic index and the pre-specified heterogeneity analysis.

Figure 1: Socio-Economic Gradient in Impacts on Behavior-Based Socio-Emotional Skills



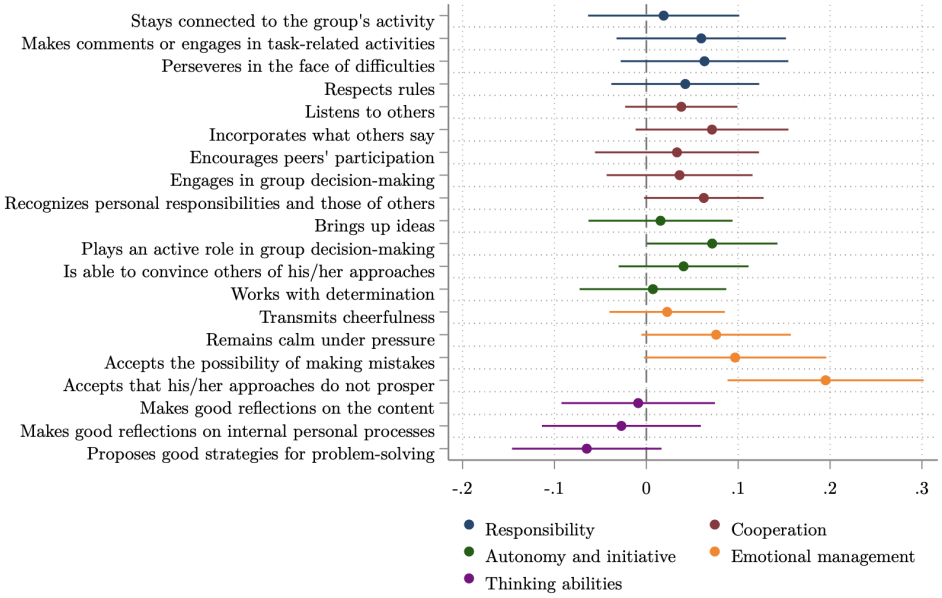
Notes: This figure plots intent-to-treat estimates of the effect of assignment to the intervention on behavior-based socio-emotional skill indices, by students' pre-treatment socio-economic background. Outcomes are standardized using the control group mean and standard deviation. Estimates are obtained from split sample regressions across different points of the index distribution. Separate series are shown for measures collected during the standardized group activity administered at endline and for in-classroom observations conducted during regular instruction. The regressions include school fixed effects, week-of-measurement fixed effects, pre-specified student socio-demographic characteristics, and baseline covariates that exhibited imbalance across treatment arms. Whiskers denote 95 percent confidence intervals. See Table E2 for the numeric values and more details on the econometric specification.

and is most pronounced for autonomy and responsibility, with additional gains in thinking abilities among disadvantaged students. Effect sizes range from approximately 0.1 to 0.5 standard deviations, depending on the domain and measurement context, and most remain statistically significant after adjustment for multiple hypothesis testing. In contrast, estimated effects for students at the upper end of the socio-economic distribution are close to zero and often negative. Appendix Table E2 reports the split-sample estimates underlying Figure 1, while Appendix Table E1 presents results using a median split of the socio-economic index.

We next examine whether treatment effects vary by students' baseline socio-emotional skills,

which are strongly correlated with socio-economic background in our sample. Appendix Tables E3 and E4 show a similar pattern when students are split by pre-treatment measures of social engagement and emotional management. Treatment effects are concentrated among students starting from weaker baseline skills. Among students with low baseline social engagement, we observe increases of approximately 0.1 to 0.36 standard deviations across nearly all socio-emotional domains and observational settings, with most estimates remaining statistically significant after correcting for multiple hypothesis testing. In contrast, no meaningful changes are detected for students with higher baseline social engagement. Results for baseline emotional management are qualitatively similar but more muted: only autonomy, responsibility, and thinking abilities show significant increases of approximately 0.27 to 0.46 standard deviations among students with lower initial emotional management skills, while effects for students with higher baseline emotional management are small and statistically indistinguishable from zero.

Figure 2: Behavior-Level Impacts Underlying Socio-Emotional Skill Domain Effects



Notes: This figure reports intent-to-treat estimates of the effect of assignment to the intervention on individual observed behaviors measured during the standardized group activity administered at endline. Each point represents a regression coefficient for a single behavior, standardized using the control group mean and standard deviation. Behaviors are grouped by socio-emotional skill domain, as indicated by color. All specifications include school fixed effects, fixed effects for the randomization strata defined by school disadvantage level, week-of-measurement fixed effects, and pre-specified student socio-demographic controls. Robust standard errors are clustered at the school-grade level. Horizontal bars denote 95 percent confidence intervals.

INDIVIDUAL BEHAVIORS: WHERE CHANGE OCCURS. To interpret the domain-level indices, Figure 2 reports treatment effects on the individual behaviors underlying the socio-emotional skill measures, as observed during the standardized group activity.

Two patterns stand out. First, within all skill domains except thinking abilities the estimated coefficients on individual behaviors are uniformly positive, though they vary in magnitude and statistical significance. This pattern indicates that the domain-level results are not driven by a single behavior, but instead reflect coordinated changes across multiple related behaviors.

Second, positive effects are concentrated in behaviors related to emotional regulation (e.g., remaining calm under pressure, accepting the possibility of making mistakes), interaction and participation (e.g., playing an active role in group decision-making, incorporating peers' contributions), and task engagement (e.g., making task-related comments or staying connected to the activity). In contrast, effects on higher-order reflective or abstract thinking behaviors are slightly negative though less precisely estimated. This mapping clarifies the nature of the domain-level gains: the intervention primarily affects how students engage with tasks, peers, and classroom demands, rather than eliciting broad changes in cognitive processing.

Taken together, the behavior-level evidence reinforces the interpretation of the earlier results. Average effects at the domain level are modest, but the intervention generates meaningful and structured behavioral changes among disadvantaged students, while producing little detectable change among more advantaged peers.

VII.2 Survey-Based Socio-Emotional Skill Measures

SELF-REPORTED SOCIO-EMOTIONAL SKILLS. Survey measures capture students' self-assessments and perceptions and therefore complement (but do not substitute for) the observer-based outcomes. Survey instruments are inherently subjective and may capture beliefs, self-concepts, or reporting styles in addition to underlying behavior. We therefore use survey-based measures primarily for triangulation and to shed light on potential mechanisms, rather than as standalone evidence of skill formation. In the main text, we focus on the five domains of the Behavioral, Emotional, and Social Skills Inventory (BESSI).

Table 2 reports impacts on the five BESSI domains. We detect a positive effect of 0.09-standard deviation increase on self-reported innovation, which captures creativity, abstract thinking, and

Table 2: Self-Reported Socio-Emotional Skills (BESSI Survey)

	(1)	(2)	(3)	(4)	(5)
	Self-Management	Social Engagement	Cooperation	Emotional Management	Innovation
Treated	0.020	0.039*	0.004	-0.004	0.094***
	(0.023)	(0.023)	(0.026)	(0.028)	(0.026)
Control mean	3.55	3.22	3.67	3.23	3.32
Observations	1892	1892	1892	1892	1892
MHT q-val.	0.65	0.24	1.00	1.00	0.00

Notes: This table reports intent-to-treat estimates of the effect of assignment to the intervention on self-reported socio-emotional skill indices measured using the 45-item Behavioral, Emotional, and Social Skills Inventory (BESSI) survey. Outcomes are standardized using the control group mean and standard deviation. All specifications include school fixed effects and fixed effects for the randomization strata defined by school disadvantage level. Week-of-measurement fixed effects are included. Regressions control for pre-specified student socio-demographic characteristics and baseline covariates that exhibited imbalance across treatment arms. Baseline measures of the outcome are also included. Robust standard errors are reported in parentheses and are clustered at the school-grade level. The row labeled "MHT q-val." reports sharpened false discovery rate q-values accounting for multiple hypothesis testing within each outcome domain. Statistical significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

information processing. There is also weaker evidence of improvements in social engagement, particularly among disadvantaged students. This is consistent with the heterogeneous improvements in cooperative and participatory behaviors observed by external raters among the same student populations. In contrast, we find no systematic changes in self-management, cooperation, or emotional management, and no detectable effects on broad personality traits or related constructs such as grit and growth mindset Appendix E).

Taken together, the survey evidence points to selective changes in self-perceived skills. Relative to the observer-based measures, survey outcomes capture a narrower set of shifts, consistent with the distinction between observable behavior and self-reported traits.

VIII Teacher Practices and Classroom Processes

Because the intervention operates through teachers and is designed to change day-to-day classroom interactions, we examine whether treatment assignment affected instructional practices and classroom processes. The outcomes in this section are not treated as primary endpoints and were not pre-specified as main paper results. We therefore interpret them as mechanism-consistent evidence that helps characterize *how* instructional attention and interaction patterns adjust in response to the program, and how those adjustments differ across classroom contexts.

VIII.1 External Observations of Teaching Practices

We use independent classroom observations to assess whether the instructional shifts implied by the intervention are visible in teachers' day-to-day practices. Observers recorded teacher–student interactions at five-minute intervals using the Teacher Roles Observation Schedule (TROS; Waxman et al., 1990). These observations were collected independently of implementation and do not track fidelity or the use of specific program tools. Instead, they provide a common, high-frequency measure of classroom interaction patterns, allowing us to assess whether teaching practices moved in directions consistent with the intervention's pedagogical logic.

The intervention does not prescribe fixed instructional scripts or mandate specific observable teaching moves. Rather, it aims to reorient classroom practice away from predominantly teacher-centered, content- and product-focused instruction toward more interactive, student-centered, and motivational approaches. We therefore interpret the observational measures as indicators of alignment with the intervention's pedagogical model, rather than as measures of compliance with particular program components.

MEASUREMENT OF INSTRUCTIONAL PRACTICES. From the TROS protocol, we construct six outcomes, each defined as the share of observation intervals in which a given practice is observed. We organize these outcomes into three dimensions that capture distinct aspects of classroom practice emphasized by the intervention: instructional organization, instructional interaction style, and regulation of learning and behavior.

Instructional organization captures how instruction is structured. We measure the share of intervals devoted to whole-class instruction, reflecting the degree to which teaching is centralized and teacher-led, as opposed to occurring in small-group, individual, or circulating formats. We also measure content- or output-focused instruction, defined as interactions centered on content delivery, product outcomes, or disciplinary exposition. Together, these measures characterize the extent to which instruction is organized around traditional, centralized delivery versus formats that allow for observation, interaction, and feedback.

Instructional interaction style captures how teachers engage with students during instruction. We measure an interactive instructional style, defined as the share of interactions involving listen-

ing, commenting, cueing or prompting, and questioning. This measure reflects a dialogic mode of instruction in which teachers actively engage with students' thinking rather than primarily delivering explanations. In addition, we measure motivational support, capturing interactions in which teachers show interest in student work, express personal regard, encourage effort, or praise behavior or performance. These measures capture the relational and feedback-oriented features of instruction emphasized by the program.

Regulation of learning and behavior captures how classroom conduct and learning are managed. We measure encouragement of student agency, defined as the share of interactions in which teachers promote self-management, questioning, extended responses, peer support, or redirection of student thinking. As a counterpoint, we also measure corrections of student behavior, capturing direct teacher interventions to address misbehavior. Considered jointly, these outcomes reflect the extent to which regulation shifts from reactive, teacher-driven control toward more proactive, student-driven self-regulation.

Table 3: Effects of the Intervention on Observed Instructional Practices

	(1)	(2)	(3)	(4)	(5)	(6)
	Instructional Organization		**Instructional interaction style**		**Regulation of learning and behavior**	
	Whole class instruction (%)	Focus on content (%)	Interactive instruct. style (%)	Motivational support (%)	Encourages student agency (%)	Corrects student behavior (%)
Treated	11.720*	-1.392	7.814**	5.033	2.704	-8.112**
	(6.378)	(5.327)	(3.462)	(4.161)	(4.478)	(3.386)
Control mean	36.44	56.37	28.77	22.76	19.46	10.26
Observations	224	224	224	224	224	224
Baseline outcome	Yes	Yes	Yes	Yes	Yes	Yes
Week-of-Survey FE	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes
MHT q-val.	0.10	0.53	0.09	0.21	0.49	0.09

Notes: This table reports intent-to-treat estimates of the effect of assignment to the intervention on observed instructional practices, measured using external classroom observations. Outcomes are defined as the share of five-minute observation intervals in which the corresponding instructional practice is recorded. Measures are grouped into three dimensions: instructional organization, instructional interaction style, and regulation of learning and behavior. All specifications include school fixed effects and week-of-measurement fixed effects. Regressions control for baseline values of the outcome. Robust standard errors are reported in parentheses and are clustered at the school-grade level. The row labeled "MHT q-val." reports sharpened false discovery rate q-values accounting for multiple hypothesis testing within each outcome domain. Statistical significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

EFFECTS ON INSTRUCTIONAL PRACTICES. Table 3 shows that assignment to the intervention led to clear changes in how instructional time was used, even though total instructional inputs were held fixed. Treated classrooms devoted a marginally larger share of time to whole-class instruction, increasing this margin by about 12 percentage points relative to a control mean of 36 percent. At the same time, teachers adopted a more dialogic instructional style: the share of interactions

characterized by listening, prompting, or questioning rose by 7.8 percentage points, corresponding to a 27 percent increase over the control mean.

The most pronounced change occurs along the regulation margin. Treated teachers spent significantly less time correcting student behavior, with a reduction of 8.1 percentage points from a control mean of just over 10 percent—an economically large decline. In contrast, we find no detectable average changes in time devoted to content- or output-focused instruction, motivational support, or explicit encouragement of student agency.

Taken together, the average effects indicate that the intervention shifted classroom interactions toward more interactive engagement and away from reactive behavioral correction, without displacing academic content delivery in the aggregate.

HETEROGENEITY BY CLASSROOM SOCIO-ECONOMIC COMPOSITION. Appendix Table E5 shows that instructional responses to the intervention differ sharply with classroom socio-economic composition. In classrooms with a high share of disadvantaged students (Panel A), treatment primarily alters the mode of instruction. Teachers substantially reduce the share of time devoted to content- or output-focused interactions, by 13.5 percentage points relative to a control mean of 61.8 percent, and simultaneously increase interactive instructional practices by 16.0 percentage points, a 55 percent increase over the control mean. In these classrooms, we detect no systematic changes in encouragement of student agency or in the correction of student behavior.

In contrast, in classrooms with a low share of disadvantaged students (Panel B), the intervention mainly affects the regulation of learning and behavior. Treated teachers increase encouragement of student agency by 13.2 percentage points (an 81 percent increase) and motivational support by 18.1 percentage points (an 83 percent increase), while sharply reducing time spent correcting student behavior by 17.5 percentage points from a control mean of 11.3 percent. In these settings, there is little evidence of changes in instructional organization or interactive instructional style.

Taken together, the heterogeneity results indicate that the same intervention induced distinct instructional reallocations depending on baseline classroom constraints: shifting instructional delivery in more disadvantaged contexts, and shifting behavioral regulation in more advantaged ones.

Table 4: Student-Reported Changes on Instructional Practices

	(1)	(2)	(3)
	Standardized Index	BART	Empirical Bayes
Treated	0.157**	0.164*	0.347*
	(0.079)	(0.086)	(0.198)
Control mean	-0.00	-0.10	-0.22
Observations	1891	1859	1891
Baseline outcome	No	No	No
Week-of-Survey FE	Yes	Yes	Yes
School FE	Yes	Yes	Yes
Socio-demographic controls	Yes	Yes	Yes
Unbalanced controls	Yes	Yes	Yes
MHT q-val.	0.09	0.09	0.09

Notes: This table reports intent-to-treat estimates of the effect of assignment to the intervention on students' perceptions of instructional practices, measured using an endline student survey. Column (1) reports results for a standardized index aggregating items on individualized teacher attention, whole-class instructional effort, instructional clarity, and peer interaction. Columns (2) and (3) report alternative index constructions using Bartlett factor scores (BART) and empirical Bayes estimates, respectively. Outcomes are standardized using the control group mean and standard deviation. All specifications include school fixed effects and week-of-survey fixed effects Regressions control for pre-specified student socio-demographic characteristics and baseline covariates that exhibited imbalance across treatment arms. Baseline measures of the outcome are not available. Robust standard errors are reported in parentheses and are clustered at the school-grade level. The row labeled "MHT q-val." reports sharpened false discovery rate q-values accounting for multiple hypothesis testing across index constructions. Statistical significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

VIII.2 Student-Perceived Changes in Instructional Practices

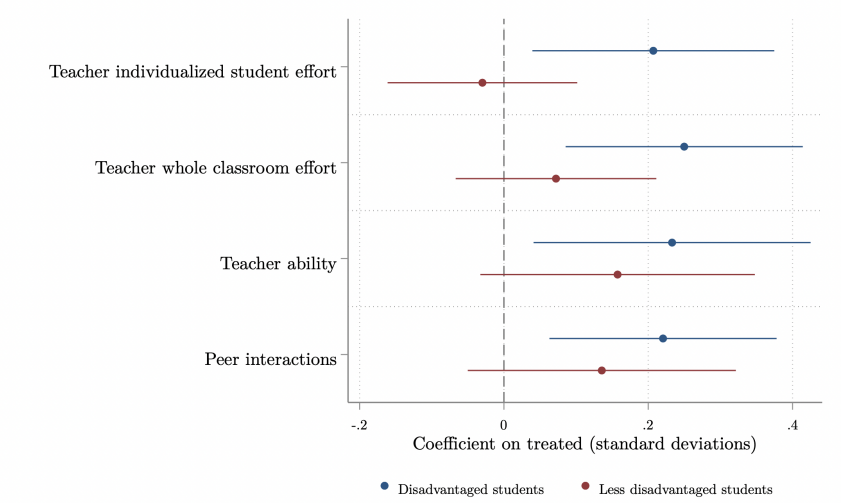
We next examine whether changes in classroom interaction patterns are salient to students. These measures capture how instructional practices are experienced by students and provide a beneficiary-facing complement to the external observation data.

Table 4 reports intent-to-treat effects on an index of student-perceived instructional practices measured at endline. The index aggregates items capturing four domains: individualized teacher attention, whole-class instructional effort, instructional clarity, and the role of peer interactions. Outcomes are standardized relative to the control group, and results are robust to alternative index constructions using Bartlett factor scores and empirical Bayes estimates.

Students assigned to treated teachers report improved instructional practices. The estimated effect on the standardized index is approximately 0.16–0.17 standard deviations across scoring methods. Figure 3 shows that these perceived changes are heterogeneous by socio-economic background: gains are concentrated among more disadvantaged students, particularly for perceived individualized attention and whole-class instructional effort. For less disadvantaged students, estimates are smaller and statistically indistinguishable from zero.

Taken together with the external observation results, the student reports suggest that the instructional adjustments induced by the program were not only observable to trained raters but also experienced by students—especially experienced among those from more disadvantaged backgrounds.

Figure 3: Student-Reported Changes in Instructional Practices



Notes: This figure plots intent-to-treat estimates of the effect of assignment to the intervention on student-reported instructional practice domains, measured at endline. Outcomes are standardized using the control group mean and standard deviation. Separate estimates are shown for students from more disadvantaged and less disadvantaged socio-economic backgrounds, based on a pre-treatment socio-economic index. Each point represents a coefficient from a regression of the domain outcome on treatment assignment, estimated separately by group. All specifications include school fixed effects and week-of-survey fixed effects. Regressions control for pre-specified student socio-demographic characteristics and baseline covariates that exhibited imbalance across treatment arms. Robust standard errors are clustered at the school-grade level. Horizontal bars denote 95 percent confidence intervals.

VIII.3 Teachers' Time Use and Classroom Management

Finally, we use teachers' self-reported time use to assess whether the program changed total workload or primarily altered within-classroom allocation. Table 5 reports intent-to-treat estimates for weekly hours spent on instructional delivery and classroom activities.

On average, we find little evidence of systematic changes in total reported in-classroom time or in most instructional activities. Estimated effects are instead concentrated in a narrow classroom management margin: treated teachers report spending less time addressing disciplinary issues. The implied reduction is economically meaningful (approximately 40 minutes per week), although it is not robust to multiple-hypothesis adjustments.

This reduction is more pronounced in more disadvantaged school settings, where baseline class-

room management demands are higher. In contrast, we do not detect consistent changes in activities related to instructional delivery formats (individual work, small group instruction, facilitating group work) in the time-use data, and we find no clear evidence of increased workload in the classroom.

Table 5: Effects on Teachers' Time Use During Classroom Instruction

	(1)	(2)	(3)	(4)	(5)	(6)
	Total	Direct whole-class instructions	Working with students individually	Small group instruction	Addressing student disciplinary issues	Facilitating group work
Treated	-2.559*	-0.066	-0.512	-1.011	-0.641**	-0.136
	(1.295)	(0.401)	(0.400)	(0.779)	(0.286)	(0.471)
Control mean	13.18	2.29	2.43	3.68	1.38	3.46
Observations	92	92	91	91	91	92
Baseline outcome	Yes	Yes	Yes	Yes	Yes	Yes
Week-of-Survey FE	Yes	Yes	Yes	Yes	Yes	Yes
School FE	Yes	Yes	Yes	Yes	Yes	Yes
MHT q-val.	0.19	0.45	0.26	0.26	0.19	0.45

Notes: This table reports intent-to-treat estimates of the effect of assignment to the intervention on teachers' self-reported weekly time allocation across in-classroom activities. Outcomes are measured in hours spent in the most recent teaching week. Each column corresponds to a distinct instructional or classroom-management activity. All specifications include school fixed effects and week-of-survey fixed effects. Regressions control for baseline measures of the outcome and pre-specified teacher characteristics. Robust standard errors are reported in parentheses and clustered at the school-grade level. The row labeled "MHT q-val." reports sharpened false discovery rate q-values accounting for multiple hypothesis testing within this outcome family. Statistical significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Overall, the self-reported time-use evidence complements the observation and student-perception results: the program appears to change the *mode* of classroom interaction and to reduce time devoted to reactive discipline, particularly in settings with higher baseline behavioral frictions. We do not interpret these estimates as causal mediators, but as descriptive evidence on which classroom margins adjust under a teacher-implemented reform.

IX Downstream Educational Outcomes

We next examine whether the intervention affected downstream educational outcomes, focusing on academic performance and school attendance. These outcomes were not directly targeted by the program and are therefore interpreted as secondary responses to changes in classroom interactions and student behavior.

Table 6 reports intent-to-treat estimates for third-term grade point average (GPA) and attendance outcomes, allowing effects to differ by students' socio-economic background. On average, we find no detectable effects on either GPA or attendance.

Despite the lack of average effects, treatment impacts differ sharply by socio-economic background. Among students from lower socio-economic backgrounds, the intervention leads to a sta-

Table 6: Effects on Academic Performance and Attendance

	(1)	(2)	(3)	(4)
	3rd term GPA		**Num. absent days**	
	Std.	Scale 0-10	Total	Unjustified
	Panel A. More underprivileged students			
Treated	-0.028	-0.033	-2.260**	-1.164
	(0.027)	(0.031)	(0.995)	(0.815)
Control mean	0.04	7.00	13.95	6.49
Observations	487	487	454	303
MHT q-val.	0.29	0.29	0.13	0.29
	Panel B. Less underprivileged students			
Treated	0.074**	0.085**	1.579**	0.656*
	(0.031)	(0.036)	(0.673)	(0.337)
Control mean	0.09	7.06	8.83	2.86
Observations	606	606	650	478
MHT q-val.	0.03	0.03	0.03	0.03
Treated × Low SES	-0.102**	-0.117**	-3.839***	-1.820*
	(0.040)	(0.046)	(0.913)	(0.975)
Control mean	0.06	7.04	11.03	4.29
Observations	1093	1093	1104	781
Baseline outcome	Yes	Yes	Yes	Yes
Socio-demographic controls	Yes	Yes	Yes	Yes
Unbalanced controls	Yes	Yes	Yes	Yes

Notes: This table reports heterogeneity in intent-to-treat estimates of the effect of assignment to the intervention on academic performance and attendance outcomes, by students' pre-treatment socio-economic background. Academic performance is measured using third-term grade point average (GPA), reported both in standardized units and on the original 0–10 grading scale. Attendance outcomes include the total number of days absent and the number of unjustified absences during the academic year. Panel A reports impacts for students with lower socio-economic background, while Panel B reports corresponding impacts for students with higher socio-economic background. The row labeled "Treated × Low SES" reports the interaction coefficient from a full-sample regression. All specifications include school fixed effects. Regressions control for baseline measures of the outcome, pre-specified student socio-demographic characteristics, and baseline covariates that exhibited imbalance across treatment arms. Robust standard errors are reported in parentheses and clustered at the school-grade level. The row labeled "MHT q-val." reports sharpened false discovery rate q-values accounting for multiple hypothesis testing within this outcome family.Statistical significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

tistically significant reduction in absenteeism. Treated low-SES students experience approximately 2.3 fewer absences during the third term, corresponding to a reduction of about 17 percent relative to the control mean, while GPA remains statistically unchanged.

In contrast, among students from higher socio-economic backgrounds, the intervention is associated with improved academic performance. GPA increases by approximately 0.075 standard deviations, with positive point estimates across the two GPA measures reported. At the same time, attendance weakly declines for this group, driven primarily by an increase in unjustified absences.

Taken together, these patterns indicate that improvements in academic performance are not

mechanically driven by increased attendance, and that the mapping from behavioral change to educational outputs differs across students. Consistent with the conceptual framework, the results suggest asymmetric translation of classroom behavior into attendance versus grades across groups. We refrain from interpreting these effects as welfare improvements and instead view them as informative about how instructional changes differentially affect educational outcomes in heterogeneous student populations.

X Discussion and Conclusion

This paper examines how a teacher-implemented intervention operates inside classrooms when instructional time, curriculum content, and formal assessment are fixed. By combining experimental variation with external observation of both teaching practices and student behavior, we trace how changes in instruction translate into behavioral responses and, only imperfectly, into academic outcomes. This approach clarifies why interventions that target socio-emotional skills often generate weak average effects on grades while producing substantial, heterogeneous changes inside classrooms.

The central empirical pattern in this paper is that student behavior responds before grades, and that these responses differ systematically across students. Observable behaviors related to autonomy, responsibility, and task engagement adjust when teachers change how they observe students, provide feedback, and structure everyday interactions. These behavioral responses are concentrated among students with weaker baseline skills and fewer external supports. For these students, improvements in classroom behavior reduce absenteeism, indicating relaxed engagement and participation constraints even when academic performance does not immediately change.

In contrast, more advantaged students exhibit little observable behavioral change but experience modest improvements in grades. This asymmetry highlights that the mapping from behavior to academic outcomes is neither uniform nor mechanical. Institutional features such as grading practices, subject-specific evaluation, and attendance rules, mediate how behavioral changes translate into formal outcomes. As a result, the same instructional change can affect different outcome margins for different students. From this perspective, the absence of average effects on grades or attendance reflects offsetting responses across groups rather than a lack of impact.

These findings help explain several recurring patterns in the literature on socio-emotional learn-

ing. First, they clarify why interventions can generate long-run gains despite limited short-run academic effects: behavioral adjustments emerge earlier and only gradually translate into conventional outcomes. Second, they show why heterogeneity is pervasive. When interventions operate by reallocating teacher attention under binding constraints, differences in baseline skills naturally generate uneven responses. Third, they illustrate why evaluations that rely solely on grades or test scores can miss welfare-relevant changes, particularly for disadvantaged students.

A key contribution of the paper is methodological. Behavior-based measures collected through external observation reveal adjustments that are largely invisible in self-reports and only partially reflected in administrative data. While such measures do not replace long-run outcomes, they provide direct evidence on where and how skill formation begins inside classrooms and allow for more precise interpretation of null or mixed results on standard metrics.

The results also inform the design and evaluation of teacher professional development. The intervention studied here does not expand instructional inputs; instead, it reshapes how teachers deploy existing effort. Making socio-emotional behavior more salient and feedback more systematic alters classroom dynamics, especially in high-need environments, without increasing workload. At the same time, the heterogeneous responses underscore that reforms based on instructional reorganization should not be expected to produce uniform gains across students or outcomes.

Although the empirical setting is education, the mechanism is broader. In many policy domains, such as healthcare, social services, and development, interventions rely on frontline agents operating under tight constraints and exercising discretion over effort allocation. In these contexts, heterogeneous behavioral responses and asymmetric outcome effects are likely to be the norm. Understanding policy effectiveness therefore requires tracing how interventions reshape behavior and attention on the ground, and for whom these changes matter.

By opening the black box of classroom processes, this paper shows that behavior is often the first margin along which meaningful change occurs, and that recognizing this margin is essential for interpreting both success and apparent failure in human capital interventions.

References

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. (2023). When Should You Adjust Standard Errors for Clustering?*. *The Quarterly Journal of Economics*, 138(1):1–35.

Akyeampong, K., Andrabi, T., Banerjee, A., Banerji, R., Dynarski, S., Glennerster, R., Grantham-McGregor, S., Muralidharan, K., Piper, B., Ruto, S., Saavedra, J., and Schmelkes, H. (2023). Cost-Effective Approaches to Improve Global Learning. What does recent evidence tell us are "Smart Buys" for improving learning in low- and middle-income countries? Technical report.

Alan, S., Boneva, T., and Ertac, S. (2019). Ever Failed, Try Again, Succeed Better: Results from a Randomized Educational Intervention on Grit. *The Quarterly Journal of Economics*, 134(3):1121–1162.

Alan, S. and Kubilay, E. (2025). Empowering Adolescents to Transform Schools: Lessons from a Behavioral Targeting. *American Economic Review*, 115(2):365–407.

Alan, S. and Mumcu, I. (2022). Nurturing Childhood Curiosity to Enhance Learning: Evidence from a Randomized Pedagogical Intervention.

Almlund, M., Duckworth, A. L., Heckman, J., and Kautz, T. (2011). Personality Psychology and Economics. *Handbook of the Economics of Education*, 4:1–181.

Anderson, M. L. (2008). 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*, 103(484):1481–1495.

Attanasio, O., Blundell, R., Conti, G., and Mason, G. (2020). Inequality in socio-emotional skills: A cross-cohort comparison. *Journal of Public Economics*, 191:104171.

Aucejo, E., Coate, P., Fruehwirth, J. C., Kelly, S., and Mozenter, Z. (2022). Teacher Effectiveness and Classroom Composition: Understanding Match Effects in the Classroom. *The Economic Journal*, 132(648):3047–3064.

Baron-Cohen, S., Wheelwright, S., Spong, A., Scahill, V., and Lawson, J. (2001). Are intuitive physics and intuitive psychology independent? A test with children with Asperger Syndrome.

Beg, S., Halim, W., Lucas, A. M., and Saif, U. (2022). Engaging Teachers with Technology Increased Achievement, Bypassing Teachers Did Not. *American Economic Journal: Economic Policy*, 14(2):61–90.

Borghans, L., Duckworth, A. L., Heckman, J. J., and ter Weel, B. (2008). The Economics and Psychology of Personality Traits. *Journal of Human Resources*, 43(4):972–1059.

Burgess, S. M., Rawal, S., and Taylor, E. S. (2022). Teachers' Use of Class Time and Student Achievement. Technical Report w30686, National Bureau of Economic Research.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., and Yagan, D. (2011). How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star. *The Quarterly Journal of Economics*, 126(4):1593–1660.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review*, 104(9):2633–2679.

Cilliers, J., Fleisch, B., Kotze, J., Mohohlwane, M., and Taylor, S. (2022). The Challenge of Sustaining Effective Teaching: Spillovers, Fade-out, and the Cost-effectiveness of Teacher Development Programs. *Economics of Education Review*, 87:102215.

Cunha, F. and Heckman, J. (2007). The Technology of Skill Formation. *American Economic Review*, 97(2):31–47.

Cunha, F., Heckman, J. J., and Schennach, S. (2010). Estimating the Technology of Cognitive and Noncognitive Skill Formation. *Econometrica*, 78(3):883–931.

Dobbie, W. and Fryer Jr., R. G. (2013). Getting beneath the Veil of Effective Schools: Evidence from New York City. *American Economic Journal: Applied Economics*, 5(4):28–60.

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

Dweck, C. S. (2000). *Self-Theories: Their Role in Motivation, Personality and Development*. Psychology Press, Philadelphia.

Dweck, C. S., Chiu, C.-y., and Hong, Y.-y. (1995). Implicit Theories and Their Role in Judgments and Reactions: A Word From Two Perspectives. *Psychological Inquiry*, 6(4):267–285.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., and Yoon, K. S. (2001). What Makes Professional Development Effective? Results from a National Sample of Teachers. *American Educational Research Journal*, 38(4):915–945.

Grimaldo, F. and López-Iñesta, E. (2021). Com classificar els centres educatius segons la complexitat? *Informes d'Avaluació 26*.

Gupta, S. (2025). Can Innovation Be Taught? Evidence from a School-Based RCT in India.

Hanushek, E. A. (2003). The Failure of Input-based Schooling Policies. *The Economic Journal*, 113(485):F64–F98.

Hargraves, V. (2019). Mindset self-evaluation for teachers.

Heckman, J. J. and Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics*, 19(4):451–464.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A., and Yavitz, A. (2010). The rate of return to the HighScope Perry Preschool Program. *Journal of Public Economics*, 94(1-2):114–128.

Heckman, J. J., Pinto, R., and Savelyev, P. A. (2013). Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes. *American Economic Review*, 103(6):2052–2086.

Heckman, J. J., Stixrud, J., and Urzua, S. (2006). The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior. *Journal of Labor Economics*, 24(3):411–482.

Huillery, E., Bouguen, A., Charpentier, A., Algan, Y., and Chevallier, C. (2025). The Role of Mindset in Education: A Large-Scale Field Experiment in Disadvantaged Schools. *The Economic Journal*, 135(671):2276–2301.

Humphries, J. E. and Kosse, F. (2017). On the interpretation of non-cognitive skills – What is being measured and why it matters. *Journal of Economic Behavior & Organization*, 136:174–185.

Jackson, C. K. (2018). What Do Test Scores Miss? The Importance of Teacher Effects on Non–Test Score Outcomes. *Journal of Political Economy*, 126(5):2072–2107.

Kautz, T., Heckman, J., Diris, R., Weel, B., and Borghans, L. (2014). Fostering and Measuring Skills: Improving Cognitive and Non-cognitive Skills to Promote Lifetime Success. *OECD Education Working Papers*, page 110.

Kern, M. L., Benson, L., Steinberg, E. A., and Steinberg, L. (2016). The EPOCH Measure of Adolescent Well-Being. *Psychological Assessment*, 28(5):586–597.

Kraft, M., Blazar, D., and Hogan, D. (2018). The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence. *Review of Educational Research*, 88:003465431875926.

Kraft, M. A. (2019). Teacher Effects on Complex Cognitive Skills and Social-Emotional Competencies. *Journal of Human Resources*, 54(1):1–36.

Lazear, E. P. (2001). Educational Production. *The Quarterly Journal of Economics*, 116(3):777–803.

Malamud, O., Mitrut, A., Pop-Eleches, C., and Urquiola, M. (2024). Self-, Peer-, and Teacher Perceptions under School Tracking.

McShane, M. (2022). How do teachers spend their time? Reporting findings from a national survey of educators in district, charter, and private schools.

Oreopoulos, P., Gibbs, C., Jensen, M., and Price, J. (2024). Teaching Teachers To Use Computer Assisted Learning Effectively: Experimental and Quasi-Experimental Evidence.

Popova, A., Evans, D. K., Breeding, M. E., and Arancibia, V. (2022). Teacher Professional Development around the World: The Gap between Evidence and Practice. *The World Bank Research Observer*, 37(1):107–136.

Rockoff, J. E. (2004). The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *American Economic Review*, 94(2):247–252.

Roth, G., Assor, A., Kanat-Maymon, Y., and Kaplan, H. (2007). Autonomous motivation for teaching: How self-determined teaching may lead to self-determined learning. *Journal of educational psychology*, 99:761.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18(2):119–144.

Santos, I., Petroska-Beska, V., Carneiro, P., Eskreis, L., Boudet, A. M. M., Berniell, I., Krekel, C., Arias, O., and Duckworth, A. L. (2024). CAN GRIT BE TAUGHT? LESSONS FROM A NATIONWIDE FIELD EXPERIMENT WITH MIDDLE-SCHOOL STUDENTS.

Sorrenti, G., Zölitz, U., Ribeaud, D., and Eisner, M. (2024). The Causal Impact of Socio-Emotional Skills Training on Educational Success. *The Review of Economic Studies*, page rdae018.

Soto, C. J. and John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, 113(1):117–143.

Soto, C. J., Napolitano, C. M., and Roberts, B. W. (2019). The Behavioral, Emotional, and Social Skills Inventory, 20-item Short Form (BESSI-20) Self-Report Form.

Soto, C. J., Napolitano, C. M., Sewell, M. N., Yoon, H. J., and Roberts, B. W. (2022). An integrative framework for conceptualizing and assessing social, emotional, and behavioral skills: The BESSI. *Journal of Personality and Social Psychology*, 123(1):192–222.

Todd, P. and Wolpin, K. I. (2018). Accounting for Mathematics Performance of High School Students in Mexico: Estimating a Coordination Game in the Classroom. *Journal of Political Economy*, 126(6):2608–2650.

Todd, P. E. and Wolpin, K. I. (2003). On the Specification and Estimation of the Production Function for Cognitive Achievement. *The Economic Journal*, 113(485):F3–F33.

Waxman, H. C., Wang, M., Lindvall, C., and Anderson, K. (1990). Teacher roles observation schedule technical manual. Technical report, Temple University, Center for Research in Human Development and Education., Philadelphia.

Online Appendix

Table of Contents

A Appendix Tables and Figures **42**

A.1 Program and Intervention Design 42

B Outcome Measurement and Index Construction **44**

B.1 Overview of Outcome Families and Data Sources 44

B.2 Behavioral Framework and Coding Scheme 44

B.3 Standardized Group Activity 47

B.4 Classroom Observation Protocols 50

B.5 Survey Instruments 52

C Sample Characteristics and Balance **54**

D Econometric Details and Robustness Analyses **57**

D.1 Additional details on the main specification 57

E Additional Results on Socio-Emotional Skill Formation and Instructional Practices **58**

F Additional Material **63**

A Appendix Tables and Figures

A.1 Program and Intervention Design

Figure A1: Project Timeline

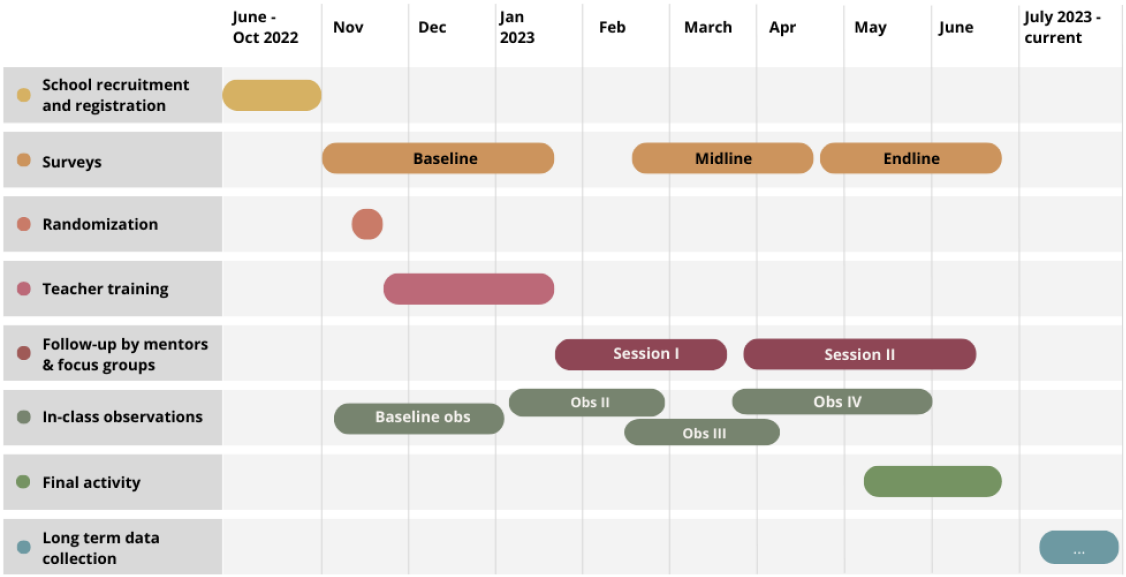


Figure A2: Educational Technologies

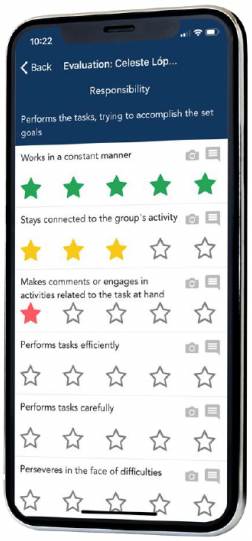


Figure A3: Snapshots of the Student Report Provided to Teachers



Skills	Observer eval.	Co-eval.	Self-eval.
Autonomy and Initiative	3.5	2	3
Cooperation	3.6	3.8	3.5
Emotion management	3.5	1.7	2.3
Responsibility	3.9	3.1	3.5
Thinking abilities	3.7	2	3.5

Behaviors with highest ratings

Cooperation	Helps peers selflessly	**4.3**	**4** Collected evidences
Cooperation	Facilitates conflict resolution	**4**	**10** Collected evidences
Responsibility	Works in a constant manner	**4**	**40** Collected evidences
Responsibility	Stays connected to the group's activity	**3.8**	**59** Collected evidences
Responsibility	Performs tasks efficiently	**3.8**	**10** Collected evidences

Behaviors with lowest ratings

Emotion management	Transmits cheerfulness	**2.7**	**3** Collected evidences
Autonomy and Initiative	Believes that he/she can initiate changes	**2.7**	**3** Collected evidences
Thinking abilities	Proposes good strategies for problem solving	**2.7**	**3** Collected evidences
Emotion management	Controls emotions when conflict arises	**2.5**	**4** Collected evidences
Cooperation	Listens to others	**2.5**	**4** Collected evidences

B Outcome Measurement and Index Construction

B.1 Overview of Outcome Families and Data Sources

This appendix provides detailed information on outcome measurement, survey instruments, behavioral coding frameworks, and index construction. Outcomes are grouped into four families: (i) student socio-emotional skills measured through behavior-based observations, (ii) survey-based student socio-emotional skills and well-being, (iii) teacher instructional practices and time use, and (iv) administrative educational outcomes. Table B1 summarizes outcome families, data sources, timing, and sample coverage.

Table B1: Outcome Families, Measurement Instruments, and Samples

Outcome family	Data source	Respondent	Timing	Sample	Role in Analysis
Student socio-emotional skills (behavior-based)	Standardized activity	External observers	EL	All students in measurement sample	Primary outcome
Student socio-emotional skills (behavior-based)	In-Classroom observations	External observers	BL / ML / EL	Random student subsample	Primary outcome
Student socio-emotional skills (survey-based)	Survey	Student (self-reported)	BL / ML / EL	All students in measurement sample	Secondary outcome (triangulation)
Teacher practices (self-reported)	Survey	Teacher	BL / EL	All treated and control teachers	Secondary outcome (mechanism)
Teacher practices (observations)	In-Classroom observations	External Observer	BL / ML / EL	All treated and control teachers	Secondary outcome (mechanism)
Grades	Administrative records	NA	BL / EL	Students with available records (non-random)	Secondary outcome (downstream)
Attendance	Administrative records	NA	BL / EL	Students with available records (non-random)	Secondary outcome (downstream)

Notes: BL = baseline; ML = midline; EL = endline. The measurement sample consists of students consented and present during data collection. Random subsamples are drawn at the classroom or student level for in-classroom observations due to resource constraints. Administrative outcomes are available only for a subset of students and are not used as primary outcomes.

B.2 Behavioral Framework and Coding Scheme

This section describes the behavioral coding framework used to construct behavior-based measures of socio-emotional skills. The intervention targets a set of 35 observable behaviors, grouped into five domains: responsibility, autonomy and initiative, cooperation, emotional management, and thinking abilities. These behaviors are designed to serve as empirical proxies for latent socio-emotional skills.

All behaviors were rated using a standardized five-point scale (1–5), with behavior-specific anchors provided to external observers. Anchors were defined in terms of concrete, observable actions rather than subjective traits, ensuring consistent interpretation across observers. A rating of

Table B2: Observable Behaviors, Measurement Context, and Behavioral Anchors (I)

Domain	Code	Behavior	Definition	Activity	Class	Missing	1-star anchor	5-star anchor
Responsibility								
Skill referring to self-discipline to achieve an objective, and to maintaining an attitude of work and effort toward the tasks to be carried out								
Responsibility	R1	Works in a constant manner	Shows sustained willingness to carry out the task	No	Yes	–	*[Anchor not available in provided anchor table]*	*[Anchor not available in provided anchor table]*
Responsibility	R2	Stays connected to the group's activity	Directs attention (what he/she looks at, what he/she listens to) toward the task. Is not distracted or carried away by other unrelated topics	Yes	Yes	No	They are not connected when they were expected to be	Completely committed to the activity at hand
Responsibility	R3	Makes comments or engages in activities related to the task at hand	Makes comments and/or takes actions when expected to do so, and the interventions are related to the task to be carried out	Yes	Yes	No	Actions and comments not related to the task	All actions and comments are oriented toward carrying out the task at hand
Responsibility	R4	Performs tasks efficiently	Does the work with the desired quality and rigor, optimizing time, materials, and energy	No	Yes	No	In carrying out the task, does not optimize time, materials, or energy	In carrying out the task, appropriately optimizes time, materials, and energy without losing quality in the work
Responsibility	R5	Performs tasks carefully	Proceeds with attention to detail, carefully and thoughtfully, in order to carry out tasks in an orderly and neat way, and thus attempts to avoid mistakes	No	Yes	No	Does not carry out tasks paying attention to details, order, neatness, etc., in order to avoid mistakes and do them as well as possible	Carries out tasks paying close attention to details, order, neatness, etc., in order to avoid mistakes and do them as well as possible
Responsibility	R6	Perseveres in the face of difficulties	Does not give up when faced with difficulties and continues working on the task, even if it does not work out on the first attempt	Yes	Yes	No	Abandons the tasks he/she is carrying out when faced with a difficulty (of any kind)	Continues working even when faced with the difficulties encountered
Responsibility	R7	Respects the rules	Respects the agreements that have been established regarding how to interact and how to use shared spaces and tools. Adapts his/her actions to the codes of practice and conduct of specific contexts	Yes	Yes	No	Disobedience of the rules and instructions that are established	Following all rules and instructions
Autonomy and Initiative								
Autonomy implies working independently, without the need for supervision, and asking for help only when required. Initiative indicates confidence to move things forward: proposing ideas and seeking solutions to problems when necessary								
Autonomy & initiative	A1	Brings up ideas	Contributes by bringing up ideas when expected to do so, thus offering new lines of thought related to the task	Yes	Yes	No	No ideas	Many ideas
Autonomy & initiative	A2	Asks questions when stuck	When faced with a difficulty that he/she cannot solve with the resources available, asks the necessary questions that allow him/her to move forward	No	Yes	Yes	Does not ask good questions that allow progress	Asks good questions in order to move forward when faced with a challenge
Autonomy & initiative	A3	Makes decisions to advance the project	Makes contributions that allow progress on the assigned task	Yes	Yes	No	Does not make contributions that allow progress	Makes contributions that allow progress
Autonomy & initiative	A4	Is able to convince others of his/her approaches	Uses ways of arguing, discussing, and presenting that convince others	Yes	Yes	Yes	Does not use forms of argumentation, discussion, or presentation that can convince others	Uses forms of argumentation, discussion, and presentation that can convince others
Autonomy & initiative	A5	Works with determination	Displays energy, enthusiasm, and commitment directed toward an activity or project. Asks questions, proposes solutions, wants to put them into practice, searches for and selects information and reads it attentively	Yes	Yes	No	Does not show enthusiasm, energy, or commitment to carrying out an activity or project	Shows the enthusiasm, energy, and commitment to carrying out an activity or project
Autonomy & initiative	A6	Believes that he/she can initiate changes	Shows self-confidence, believes that he/she can influence outcomes, and feels capable of learning and improving in order to perform the task well	No	Yes	No	Does not have the necessary confidence to influence decisions (if working in a group) or in his/her own individual work, and therefore does not carry them out. Does not feel capable of learning or improving	Has a lot of confidence in his/her contributions and makes them satisfactorily. Feels capable of learning and improving
Cooperation								
Skill that includes behaviors that facilitate group work: actions related to participation, active listening, shared decision-making, and conflict resolution								
Cooperation	C1	Listens to others	Listens actively when others speak, avoids distractions, and maintains attention	Yes	Yes	No	Does not listen actively when others speak, ignoring distractions and maintaining attention	Listens actively when others speak, ignoring distractions and maintaining attention
Cooperation	C2	Incorporates what others say	Listens actively to others and acknowledges what they say in order to add to, challenge, modify, or simply accept the idea or comment. All of this without passing judgment on the person	Yes	Yes	No	Does not listen to others and recognize what they say in order to add to, debate, modify, or accept the idea/comment. Does not imply judgment	Listens to others and recognizes what they say in order to add to, debate, modify, or accept the idea/comment. Does not imply judgment
Cooperation	C3	Encourages peers' participation	Promotes participation by creating spaces and tools, addressing others (actively), or allowing them to speak (passively)	Yes	Yes	No	Does not allow others to participate. Too autonomous	Allows others to participate and creates the necessary spaces for that to happen
Cooperation	C4	Engages in group decision making	Shares his/her comments and participates in tasks in a way that facilitates the group making decisions that move the task forward (contributes ideas, summarizes and organizes others' ideas, proposes strategies)	Yes	Yes	No	Does not participate in group decision-making	Puts forward ideas, syntheses, and organizes others' ideas, proposing strategies, etc.
Cooperation	C5	Facilitates conflict resolution	Shows skills to de-escalate an existing conflict through humor, listening, and/or patience, contributes different perspectives, etc. Does not provoke conflict.	No	Yes	Yes	Does nothing to resolve the conflict, and passivity may even allow it to continue growing	Shows skills to de-escalate an existing conflict through humor, listening, or patience, providing new perspectives. Does not generate conflict
Cooperation	C6	Recognizes personal responsibilities and those of others	Exercises his/her responsibility and promotes, respects, and accepts the value of the work and responsibilities of others	Yes	Yes	No	Does not do his/her part and does not recognize the responsibilities of others in carrying out the task at hand	Shows that he/she assumes responsibility and also recognizes the responsibilities of others within the group
Cooperation	C7	Helps peers selflessly	Shows attention, solidarity, and empathy toward others beyond the work they have to do and without expecting anything in return	No	Yes	Yes	Does NOT show attention, solidarity, or empathy toward others beyond the work they have to do and without expecting anything in return	Shows attention, solidarity, and empathy toward others beyond the work they have to do and without expecting anything in return

Notes: "Activity" indicates behaviors rated during the standardized group activity at endline; "Class" indicates behaviors rated during in-classroom observations. "Missing" reports whether missing values were allowed for that behavior in the standardized activity evaluations. The definitions of the skill domain and behaviors come from the pedagogical kit provided as part of the intervention. The anchor text comes from the external observers' training documentation.

Table B3: Observable Behaviors, Measurement Context, and Behavioral Anchors (II)

Domain	Code	Behavior	Definition	Activity	Class	Missing	1-star anchor	5-star anchor
Emotional Management								
Skill that involves being aware of one's own emotions and those of others, and managing them appropriately								
Emotional management	E1	Transmits cheerfulness	Creates positive environments, lifts spirits, and encourages others through good humor and personal enjoyment. Does not appear angry or sullen	Yes	Yes	No	Shows a negative attitude and does not contribute to others	Shows a positive and constructive attitude, raises the morale of the group
Emotional management	E2	Remains calm under pressure	Maintains calm and serenity in stressful situations. Does not show anxiety or become blocked when faced with a concern	Yes	Yes	No	Becomes anxious/blocked under pressure	Remains calm under pressure
Emotional management	E3	Controls emotions when conflict arises	Regulates emotions associated with conflict (anger, frustration, fear, shame, rage) in a way that is not disruptive, aggressive, or violent, and contributes to an improvement of the situation for oneself or for others	No	Yes	Yes	Is not able to regulate anger, frustration, fear, shame, or rage, giving rise to disruptive situations during the activity	Regulates the emotions associated with conflict (anger, frustration, fear, shame, rage) in a way that is not disruptive, aggressive, or violent, contributing to an improvement of the situation for him/herself or for others
Emotional management	E4	Accepts the possibility of making mistakes	Knows how to manage frustration and does not become irritated, angry, paralyzed, or discouraged when facing a mistake. Conceives error as an opportunity for learning	Yes	Yes	Yes	Becomes frustrated and angry when he/she makes a mistake	Shows a growth mindset when he/she makes a mistake
Emotional management	E5	Accepts that his/her approaches do not prosper	Does not become irritated, angry, or discouraged when his/her proposals are not carried out or accepted. Shows flexibility	Yes	Yes	Yes	Becomes really frustrated when his/her ideas do not prosper	Shows flexibility when his/her proposals do not advance
Emotional management	E6	Adapts behavior to the circumstances	Shows understanding of the environment and of others' emotions and adapts his/her behavior to respond and/or contribute in a positive way	No	Yes	No	Does not have a good understanding of his/her environment, and therefore does not contribute	Shows understanding of the environment and of others' emotions, and adapts his/her behavior to respond and/or contribute in a positive way
Thinking Abilities								
Thinking abilities make it possible to capture information and process it in order to build and organize knowledge, so that it can be used to solve problems in different situations. They involve intellectual inquiry, idea generation, and metacognitive skills								
Thinking abilities	T1	Relates new content with previous knowledge	Connects new content with other content already known from past topics, other disciplines, or previous experiences, with the aim of understanding it better	No	Yes	No	Does not make content connections	Connects new content with other content already known from past topics, other disciplines, or from previous experiences, with the aim of understanding it better
Thinking abilities	T2	Makes good reflections on the content covered	Demonstrates understanding of the content covered beyond merely listing or recognizing it. Is able to analyze, evaluate, express, verify, and contrast arguments related to the content addressed	Yes	Yes	No	Reflection of low/poor quality	Reflection of good quality
Thinking abilities	T3	Makes good reflections on internal personal processes	Shows awareness of how he/she feels, thinks, proceeds, and reacts. Is able to express both his/her learning processes and emotional processes	Yes	Yes	No	Reflection of low/poor quality	Reflection of good quality
Thinking abilities	T4	Asks good questions (to improve understanding or to move forward)	Asks questions that show curiosity and interest in understanding, learning, or advancing comprehension	No	Yes	No	Does not ask good questions	Asks questions that show curiosity and interest in understanding, learning, or advancing in comprehension
Thinking abilities	T5	Has creative ideas (explores alternative paths)	Produces original ideas and finds alternative paths to solve activities or problems. These ideas constitute positive and useful contributions	No	Yes	No	Does not contribute ideas that are very novel and provide a differential value	Produces original ideas and finds alternative paths to solve activities or problems. These ideas constitute positive and useful contributions
Thinking abilities	T6	Proposes good strategies for problem-solving	Understands the problem, extracts relevant information, organizes it, analyzes it, and uses it to solve the problem	Yes	Yes	Yes	Strategy of low/poor quality	Strategy of good quality
Thinking abilities	T7	Plans and prioritizes tasks	Identifies and orders tasks according to their relevance. Defines the timing of the actions to be undertaken, taking into account resources, available knowledge, and task priorities	No	Yes	No	Does not plan or prioritize	Identifies and orders tasks according to their relevance. Defines the timing of the actions to be undertaken, taking into account resources, available knowledge, and the priorities of the task
Thinking abilities	T8	Willing to discover different perspectives	Shows openness to examining other perspectives or points of view in a rational manner. Does not show rigidity, intolerance, prejudice, or inflexibility	No	Yes	No	Shows rigidity and inflexibility when finding solutions	Shows him/herself open to examining other perspectives or points of view rationally. Does not show rigidity, intolerance, prejudice, or inflexibility
Thinking abilities	T9	Expresses ideas effectively (correction, precision and structure)	Expresses him/herself in a correct, precise, and structured way. This expression may take place through different media: graphical, oral, written, conceptual maps, mathematical, etc.	No	Yes	No	Poor expression	Expresses him/herself in a correct, precise, and structured way. This expression can occur through different media: graphical, oral, written, concept maps, mathematical...

Notes: "Activity" indicates behaviors rated during the standardized group activity at endline; "Class" indicates behaviors rated during in-classroom observations. "Missing" reports whether missing values were allowed for that behavior in the standardized activity evaluations. The definitions of the skill domain and behaviors come from the pedagogical kit provided as part of the intervention. The anchor text comes from the external observers' training documentation.

1 corresponds to the absence of the targeted behavior or the presence of behavior inconsistent with task engagement, while a rating of 5 corresponds to sustained, high-quality manifestations of the behavior. Intermediate values capture partial or inconsistent expressions.

For each behavior, observers were provided with detailed guidance specifying (i) the type of actions or verbalizations corresponding to low and high ratings, (ii) whether non-manifestation of the behavior should be coded as a low rating or as missing, and (iii) how to treat edge cases such as

passive non-participation. Missing values were allowed only for a limited subset of behaviors for which non-manifestation could not be meaningfully interpreted.

All behavioral observations were conducted by external observers who were not involved in the design or implementation of the intervention. Observers were blind to treatment assignment and were not informed whether the classrooms they visited belonged to the treatment or control group. During data collection, observers did not interact with teachers or students beyond what was strictly necessary to carry out the observation protocols. Prior to fieldwork, observers received structured training on the behavioral framework and coding scheme, including detailed review of the behavior definitions and rating anchors, calibration exercises using sample videos and vignettes, mock coding sessions in out-of-the-study classrooms, and consensus discussions to align interpretations of edge cases. This training aimed to ensure consistent application of the coding rules across observers and to minimize subjective variation in ratings.

The full list of behaviors, domain classifications, rating anchors, and missing-data rules is reported in Tables B2 and B3.

B.3 Standardized Group Activity

At endline, students in the measurement sample participated in a standardized, one-hour group activity designed to elicit socio-emotional behaviors in a common environment. The activity was constructed to generate observable variation in cooperation, autonomy and initiative, emotional management, responsibility, and thinking abilities under time constraints, shared incentives, and structured reflection.

The standardized activity complements in-classroom observations by placing students in a novel, non-instructional setting that requires coordination, planning, self-regulation, and peer interaction, reducing confounding from teacher-specific practices, classroom routines, or subject content.

ACTIVITY STRUCTURE. The activity followed a uniform, pre-specified protocol implemented identically across treatment and control groups.

Students were randomly assigned to small working groups (typically 4–6 students per group, depending on class size and attendance) using a pre-generated random assignment list prepared by

the field team. Groups were asked to complete a collaborative construction task under explicit time, material, and rule constraints. The task required building a free-standing structure using a limited set of materials (e.g., straws, sticks, elastic bands, interlocking bricks) to suspend objects above the table, with performance determined by both outputs and resource use.

The activity consisted of four sequential phases:

1. *Instructions and Group Formation (15 minutes)*: A trained facilitator explained the task rules, scoring system, and constraints to the entire class. Instructions were delivered publicly to ensure that all students received identical information. Teachers were present but instructed not to intervene except when strictly necessary for classroom management or special educational needs.

2. *First construction phase (10 minutes):* Groups worked collaboratively to construct a structure using a limited set of materials. Points were awarded based on performance outcomes and deducted based on material usage and rule violations. During this phase, students were required to coordinate roles, make decisions, manage time and resources, and respond to emerging difficulties.

3. *Individual reflection and group discussion (15 minutes):* Students completed a short individual written reflection addressing their own experience, perceived group dynamics, and potential strategies for improvement. This was followed by a brief group discussion, allowing students to exchange perspectives and reassess their approach.

4. *Second construction phase (5 minutes):* Groups were given an unannounced additional period to revise and improve their structure based on their reflections, allowing observers to capture behavioral adjustments following feedback and reflection.

The session concluded with performance assessment and the completion of self- and peer-assessments aligned with the behavioral framework.

BEHAVIORAL OBSERVATION AND MEASUREMENT. Each working group was assigned a single external observer, responsible for observing all students in the group throughout the activity.

Observers rated behavior using a tablet-based application and a pre-defined subset of 20 observable behaviors drawn from the full behavioral framework described in Appendix B.2.

Observers were instructed to position themselves to maximize visibility while minimizing interaction with students. They were permitted to respond only to brief clarification questions related to task instructions and were prohibited from guiding student behavior or strategy.

Behavioral ratings were recorded at multiple points during the activity, covering early coordination and engagement, cooperation and decision-making during construction, emotional regulation under time pressure, and reflective reasoning during the reflection phase. Ratings followed the same anchored five-point scale used in classroom observations.

In addition to observer ratings, students completed brief self-assessments and peer assessments aligned with the same behavioral dimensions. For peer assessments, each student evaluated a randomly assigned subset of their group members. These data are used to construct complementary measures of perceived skills and skill awareness.

OBSERVER TRAINING AND BLINDING. All observers were external to the schools and had no prior relationship with participating teachers or students. Observers were blind to treatment assignment and were not informed whether the classroom they observed belonged to the treatment or control group.

Observer training included detailed instruction on the behavioral framework and coding rules, calibration using anchor examples, mock observation exercises, and consensus discussions to harmonize interpretation of ambiguous cases. Observers did not receive any information about classroom-level implementation of the intervention and did not interact with teachers beyond logistical coordination.

The standardized activity was administered uniformly across all schools following the same written protocol. All materials, timing, instructions, and scoring rules were standardized. Deviations (e.g., student absences, consent restrictions) were recorded on structured field logs.

Students without parental consent participated in the activity but were excluded from behavioral observation and digital data collection, in accordance with ethical requirements.

POST-ACTIVITY SELF-, PEER-, AND OBSERVER ASSESSMENTS. Following the standardized group activity, students and external observers completed a brief summative survey assessing socio-emotional behaviors manifested during the task. The external observer who rated behaviors during the activity also completed the same survey for each student in the assigned group, allowing the construction of parallel measures from self, peer, and observer perspectives. For both students and observers, responses were optional, and respondents could select "I don't know / I can't rate" when a behavior could not be meaningfully assessed.

Observers additionally reported whether each completed assessment was reliable, based on the student's behavior and engagement during survey completion (e.g., distraction, misbehavior). This information is used to flag potentially unreliable responses in robustness checks.

The summative survey was designed to provide a concise snapshot of students' socio-emotional behaviors while minimizing respondent burden and avoiding explicit references to the intervention's skill labels. The instrument consists of 10 items capturing the five behavioral domains targeted by the intervention. These items are a subset of behaviors drawn from the BESSI (192 items, Soto et al., 2022), selected through an iterative mapping and triangulation process linking (i) the 35 intervention behaviors, (ii) the BESSI item pool, and (iii) behaviors feasibly elicited by the standardized activity design.

The selection and wording of the 10 items were reviewed and validated by an external expert in education and socio-emotional learning to ensure conceptual alignment and clarity.To facilitate consistent interpretation, extreme behavioral anchors (corresponding to ratings of 1 and 5) were developed for each item. Data from the summative survey are used to construct measures of perceived skills and skill awareness by comparing self-, peer-, and observer assessments. The full list of items and anchors is reported in Table B4.

B.4 Classroom Observation Protocols

Repeated classroom observations were conducted during regular instruction between January and May 2023 by trained external observers hired and managed by an independent organization (*Empieza por Educar*). Observers had no prior relationship with participating schools, teachers, or students. Each participating classroom was visited between four and six times over the academic year, symmetrically across treatment and control groups.

Table B4: Summative Survey Items and Anchoring Examples

Survey Item	1-point Anchor (Low)	5-point Anchor (High)
Kept myself from getting distracted	I got distracted and wasted my time.	I was focused and made the most of my time.
Worked hard to succeed	I didn't put any effort into the task.	I put a lot of effort carrying out the task.
Worked with people towards a shared goal	We did not agree and have not been able to work well together.	We have cooperated and worked very well together.
Stayed calm in stressful situations	I became very agitated in stressful situations.	I remained calm in stressful situations.
Explained what I am thinking and feeling	I did not express what I felt and thought at any moment.	I did express what I felt and thought at all times.
Followed the rules	In many occasions I did not follow the rules.	I followed the rules at all times.
Changed people's mind	My proposals were never accepted by my team.	My proposals were always accepted by my team.
Have taken another person's perspective	I did not see or consider how others felt at any point.	I have seen and considered how others feel at all times.
Took responsibility when I made a mistake	When I made a mistake I always made excuses or blamed others and did not try to fix it.	When I made a mistake I always accepted it and tried to fix it.
Found logical solutions to problems	I was not able to find good solutions to problems.	I was able to find good solutions to problems.

Notes: Items were answered on a five-point Likert-type scale. Anchoring examples for the lowest (1) and highest (5) response categories were provided to students to facilitate consistent interpretation of the scale. Intermediate values (2–4) reflect partial or mixed manifestations of the behavior.

Classroom observations focused on two complementary dimensions. First, observers evaluated the manifestation of the 35 observable student behaviors targeted by the intervention, as defined in the behavioral coding framework described in Appendix B.2. These observations captured socio-emotional behaviors as they arose during routine instructional activities in a natural classroom setting. Second, at baseline and endline, observers also assessed teacher roles and teacher–student interactions using the Teacher Roles Observation Schedule (TROS; Waxman et al., 1990), a standardized classroom observation instrument.

Each classroom was assigned two external observers. During each observation session, observers independently rated student behavior using a tablet-based application and a standardized observation protocol. Within each session, the two observers evaluated a common subset of two students, while each observer additionally rated four distinct students drawn from a randomly selected pool of eight to ten students per classroom.[13] This planned overlap was designed to promote consistency in the application of behavioral coding rules while maintaining broad coverage of students within each classroom.

Observers recorded behavior ratings in real time and followed identical procedures across treatment and control classrooms. Observation timing, student selection, rating scales, and coding rules were fully standardized. Observers were blind to treatment assignment and did not interact with

[13] A ranked list of replacement students was provided in case any assigned students were absent.

teachers beyond what was strictly necessary for logistical coordination.

While formal inter-rater reliability statistics are not reported, the combination of structured training, standardized coding protocols, and planned observer overlap was intended to promote consistency and reduce subjective variation in behavioral ratings across observers.

B.5 Survey Instruments

This appendix describes the survey instruments used to measure socio-emotional skills, well-being, perceptions, and instructional practices for students and teachers. Surveys complement behavior-based measures by providing validated self-, peer-, and teacher-reported assessments. They are not used as primary outcomes unless explicitly stated.

All surveys were administered by an external survey firm following standardized protocols. Student surveys were administered in classrooms using tablets. Teacher surveys were administered in person at baseline and electronically at endline. Survey procedures were identical across treatment and control groups. Surveys were administered in either Catalan or Spanish, according to each respondent's self-reported preferred survey language.

STUDENT SURVEY INSTRUMENTS. Student surveys collected information on demographics and socio-economic background, classroom friendship networks, well-being, and socio-emotional skills.

Well-being was measured using the EPOCH Measure of Adolescent Well-Being (Kern et al., 2016). Information on parental relationships was collected using items adapted from Soto et al. (2022). Classroom friendship networks were elicited using standard name-generator questions.

Socio-emotional skills were measured using multiple validated instruments. At baseline, students completed the 45-item Behavioral, Emotional, and Social Skills Inventory (BESSI-45; Soto et al., 2022), which measures five broad socio-emotional skill domains: self-management, social engagement, cooperation, emotional resilience, and innovation skills. Students also completed the 30-item Big Five Inventory–2 Short Form (BFI–2–S; Soto and John, 2017), which measures the five major personality dimensions: agreeableness, extraversion, conscientiousness, openness to experience, and negative emotionality (neuroticism).

At endline, to reduce respondent burden, shorter versions were administered: the BESSI-20

(Soto et al., 2019) and the 15-item BFI–2 Extra-Short Form (BFI–2–XS; Soto and John, 2017). These short forms capture the same latent dimensions as the longer instruments with slightly reduced precision.

Social intelligence was measured using the Reading the Mind in the Eyes Test (RMET; Baron-Cohen et al., 2001), which assesses the ability to infer others' mental and emotional states. At endline, students additionally completed the 3-item growth mindset scale (Dweck et al., 1995; Dweck, 2000) and grit scale (Short Grit Scale; Duckworth and and Quinn, 2009).

PEER AND TEACHER ASSESSMENTS OF STUDENT SKILLS. At midline and endline, students provided peer assessments of socio-emotional skills for a randomly assigned subset of classmates using the BESSI-20 questionnaire. Random assignment of peers to be rated limits scope for selective evaluation.

At endline, teachers assessed the socio-emotional skills of a randomly selected subset of students within each classroom using the BESSI-20 Observer Short Form (Soto et al., 2019, 2022). The students evaluated by teachers correspond to the same subset selected for in-class behavioral observations.

TEACHER SURVEY INSTRUMENTS. Teacher surveys collected information on demographics, teaching experience, socio-emotional skills, instructional practices, time use, motivation, and teaching mindset.

Socio-emotional skills were measured using the BESSI-45 and BFI–2–S instruments. Time allocation within and outside the classroom was measured using the Teacher Time-Use Questionnaire (McShane, 2022). Teaching motivation was measured using a scale of autonomous motivation for teaching (Roth et al., 2007). Teaching mindset was measured using items adapted from growth mindset scales (Dweck et al., 1995; Dweck, 2000) and selected items from the Mindset for Teaching questionnaire (Hargraves, 2019).

Baseline teacher surveys were administered in person; endline surveys were administered electronically. Survey content and administration procedures were identical across treatment and control groups.

C Sample Characteristics and Balance

Balance is assessed by regressing each baseline covariate on the treatment indicator and testing for differences across treatment arms. In addition, we conduct joint tests of balance by regressing treatment assignment on the full set of baseline covariates and testing their joint significance. Stratification-cell fixed effects are included in all balance regressions. For individual covariate tests, observations with missing baseline values are excluded. For joint tests, missing values are set to zero and accompanied by missingness indicators.

Table C1: Balance on Grade Characteristics (at the student level)

	(1)	(2)	(3)	(4)
	Full Sample	Control	Treatment	Mean Difference
Panel A: Program registration information				
Number of teachers by grade	2.29	2.34	2.23	-0.13
	(0.99)	(0.95)	(1.03)	(0.22)
Number of subjects by grade	3.27	3.29	3.25	-0.04
	(1.61)	(1.46)	(1.76)	(0.38)
Number of hours by grade	17.51	17.64	17.37	-0.53
	(14.47)	(12.16)	(16.54)	(3.32)
Observations	2068	1053	1015	
Joint significance test				0.881
Panel B: School enrollment				
Total number of students	278.56	278.60	278.53	-2.47
	(120.09)	(124.24)	(115.67)	(28.86)
Total number of male students	141.82	142.17	141.46	-1.80
	(60.91)	(63.55)	(58.06)	(14.80)
Total number of female students	136.74	136.43	137.07	-0.68
	(60.81)	(62.27)	(59.27)	(14.34)
Num. Students 1 ESO: total	75.65	76.53	74.74	-2.45
	(29.01)	(30.49)	(27.36)	(6.69)
Num. groups in 1 ESO	2.70	2.72	2.69	-0.05
	(0.98)	(1.02)	(0.94)	(0.23)
Num. Students 2 ESO: total	76.84	76.62	77.06	-1.03
	(27.29)	(28.89)	(25.53)	(6.40)
Num. groups in 2 ESO	2.68	2.66	2.71	0.00
	(0.92)	(0.95)	(0.89)	(0.22)
Num. Students 3 ESO: total	73.06	72.88	73.26	-1.03
	(26.08)	(27.65)	(24.35)	(6.09)
Num. groups in 3 ESO	2.61	2.60	2.62	-0.03
	(0.90)	(0.95)	(0.86)	(0.22)
Num. Students 4 ESO: total	73.91	73.54	74.29	-0.01
	(24.40)	(25.37)	(23.37)	(6.05)
Num. groups in 4 ESO	2.56	2.56	2.56	-0.03
	(0.84)	(0.87)	(0.81)	(0.21)
Observations	2068	1053	1015	
Joint significance test				0.998
Panel C: Neighborhood socio-economic characteristics				
Socioeconomic Index at the Municipality level	96.38	96.59	96.16	-0.65
	(9.85)	(9.44)	(10.26)	(2.19)
Income source: other subsidies, per person, by section	776.89	790.76	762.51	-26.75
	(158.66)	(160.73)	(155.26)	(35.88)
Income source: other income, per person by section	1473	1454	1493	26
	(1043)	(1065)	(1019)	(223)
Income source: pensions, per person, by section	3260	3250	3271	-6
	(1085)	(1029)	(1141)	(249)
Income source: unemployment subsidies, per person, by section	784.55	786.61	782.41	-1.35
	(135.87)	(130.83)	(140.94)	(32.61)
Income source: wage, per person	9046	9023	9070	12
	(2490)	(2528)	(2452)	(610)
Income distribution P80/P20, by section	2.73	2.72	2.73	0.02
	(0.39)	(0.39)	(0.38)	(0.09)
Gini index, by section	30.47	30.43	30.51	0.11
	(3.32)	(3.43)	(3.20)	(0.75)
Average age, by section	42.81	42.82	42.80	-0.08
	(3.06)	(2.90)	(3.21)	(0.70)
Population, by section	1670	1660	1680	29
	(619)	(592)	(645)	(144)
Share of individual housholds, by section	27.95	28.04	27.86	-0.20
	(5.53)	(5.72)	(5.32)	(1.33)
Share of population aged 65 or older, by section	19.41	19.34	19.48	-0.00
	(5.33)	(5.02)	(5.64)	(1.21)
Share of Spanish population	81.43	81.54	81.31	-0.44
	(10.29)	(10.46)	(10.12)	(2.52)
Share of population aged under 18 y/o, by section	17.41	17.30	17.52	0.23
	(3.23)	(3.15)	(3.31)	(0.75)
Average houshold size, by section	2.59	2.58	2.61	0.04
	(0.49)	(0.49)	(0.49)	(0.11)
Mean income per consumption unit, by section	19029	18978	19081	21
	(4097)	(4067)	(4129)	(974)
Median income per consumption unit, by section	17284	17238	17331	18
	(3664)	(3701)	(3626)	(893)
Mean gross income per houshold, by section	39529	39367	39696	172
	(9683)	(9572)	(9800)	(2345)
Mean gross income per person, by section	15341	15304	15378	4
	(3754)	(3758)	(3751)	(884)
Mean net income per houshold, by section	32921	32810	33036	115
	(6910)	(6844)	(6979)	(1686)
Mean net income per person, by section	12774	12751	12799	-4
	(2685)	(2692)	(2678)	(636)
Observations	2068	1053	1015	
Joint significance test				1.000

Notes: Robust standard errors are shown in parentheses and clustered at the school-grade level. Statistical significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table C2: Balance on Student-Level Characteristics

	(1)	(2)	(3)	(4)
	Full sample	Control	Treatment	Mean difference
Panel A: Pre-specified covariates to be included as socio-economic controls				
Household socio-economic index	6.55	6.53	6.56	0.02
	(1.01)	(1.03)	(0.99)	(0.05)
Gender: Male	0.51	0.50	0.52	0.02
	(0.50)	(0.50)	(0.50)	(0.02)
Gender: Female	0.47	0.47	0.47	-0.00
	(0.50)	(0.50)	(0.50)	(0.02)
Gender: Other	0.02	0.03	0.01	-0.01**
	(0.14)	(0.16)	(0.11)	(0.01)
Caregivers' country: HIC, non-Spanish speaking	0.02	0.03	0.02	-0.01
	(0.15)	(0.16)	(0.14)	(0.01)
Caregivers' country: HIC, Spanish speaking	0.00	0.00	0.00	0.00
	(0.06)	(0.05)	(0.06)	(0.00)
Caregivers' country: UMIC, non-Spanish speaking	0.03	0.03	0.03	-0.00
	(0.17)	(0.18)	(0.17)	(0.01)
Caregivers' country: UMIC, Spanish speaking	0.09	0.08	0.09	0.01
	(0.28)	(0.28)	(0.29)	(0.02)
Caregivers' country: LMIC, non-Spanish speaking	0.14	0.13	0.14	0.01
	(0.34)	(0.34)	(0.35)	(0.02)
Caregivers' country: LMIC, Spanish-speaking	0.02	0.03	0.02	-0.01
	(0.15)	(0.16)	(0.14)	(0.01)
Caregivers' country: LIC, non-Spanish speaking	0.01	0.01	0.01	0.00
	(0.09)	(0.09)	(0.09)	(0.00)
Caregivers' country: Spain	0.61	0.62	0.60	-0.02
	(0.49)	(0.49)	(0.49)	(0.04)
Caregivers' country: Unknown	0.08	0.07	0.09	0.01
	(0.27)	(0.26)	(0.28)	(0.02)
Observations	2068	1053	1015	
Joint significance test				0.720
Panel B: All other socio-economic characteristics				
Number of people in the household	4.14	4.12	4.17	0.06
	(1.01)	(1.02)	(1.01)	(0.05)
Household: How many bedrooms do you have in your house?	3.04	3.04	3.04	0.00
	(0.66)	(0.67)	(0.65)	(0.04)
Household: Do you have heater in your house?	0.90	0.90	0.91	0.01
	(0.30)	(0.30)	(0.29)	(0.01)
Household: Do you have Wifi (internet) in your house?	0.99	0.99	0.98	-0.01
	(0.11)	(0.10)	(0.12)	(0.01)
Household: Do you have a computer in your house?	0.92	0.92	0.92	0.01
	(0.27)	(0.27)	(0.26)	(0.02)
Household: Do you have a tablet in your house?	0.69	0.69	0.70	0.01
	(0.46)	(0.46)	(0.46)	(0.03)
Catalan as preferred language	0.46	0.44	0.47	0.02
	(0.50)	(0.50)	(0.50)	(0.04)
Student born in Spain	0.87	0.88	0.87	-0.01
	(0.33)	(0.33)	(0.34)	(0.02)
Caregiver 1: Born in Spain	0.62	0.62	0.61	-0.02
	(0.49)	(0.49)	(0.49)	(0.04)
Caregiver 2: Born in Spain	0.64	0.65	0.63	-0.02
	(0.48)	(0.48)	(0.48)	(0.04)
Caregiver 1: No studies	0.03	0.01	0.05	0.03***
	(0.17)	(0.12)	(0.22)	(0.01)
Caregiver 1: Primary education	0.06	0.07	0.06	-0.01
	(0.24)	(0.25)	(0.24)	(0.01)
Caregiver 1: Secondary education	0.17	0.18	0.17	-0.01
	(0.38)	(0.38)	(0.38)	(0.02)
Caregiver 1: Professional education	0.16	0.16	0.17	0.01
	(0.37)	(0.37)	(0.38)	(0.02)
Caregiver 1: Attended university	0.42	0.42	0.41	-0.02
	(0.49)	(0.49)	(0.49)	(0.02)
Caregiver 2: No studies	0.03	0.02	0.04	0.02*
	(0.18)	(0.16)	(0.20)	(0.01)
Caregiver 2: Primary education	0.08	0.07	0.09	0.01
	(0.27)	(0.26)	(0.29)	(0.01)
Caregiver 2: Secondary education	0.21	0.23	0.18	-0.06**
	(0.41)	(0.42)	(0.38)	(0.03)
Caregiver 2: Professional education	0.19	0.18	0.20	0.02
	(0.39)	(0.38)	(0.40)	(0.02)
Caregiver 2: Attended university	0.34	0.34	0.34	-0.00
	(0.48)	(0.48)	(0.47)	(0.03)
Observations	2068	1053	1015	
Joint significance test				0.048**

Notes: Robust standard errors are shown in parentheses and clustered at the school-grade level. Statistical significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

D Econometric Details and Robustness Analyses

This appendix provides supplementary econometric details and robustness analyses for the results presented in Section VI. The main results are based on a fully saturated ANCOVA specification that respects the experimental design and includes pre-specified covariates and fixed effects. The analyses reported here assess the sensitivity of the estimates to alternative specifications, data-driven covariate selection, and consent and attrition patterns.

D.1 Additional details on the main specification

This subsection clarifies implementation details of the baseline ANCOVA specification used throughout the paper. The purpose is to document how the estimating equation is operationalized, not to introduce alternative models.

The baseline specification includes school fixed effects to absorb all time-invariant differences across schools, fixed effects for the randomization strata defined by school disadvantage level and, when outcomes are collected over extended fieldwork periods, fixed effects for the week of measurement to account for survey timing and seasonality.

For outcomes based on external observer ratings, the main specification additionally includes observer fixed effects to absorb systematic differences in rating styles across observers. These fixed effects are included in all reported specifications for observer-based outcomes.

The individual-level covariates included in the baseline specification are pre-specified and consist of student gender, indicators classifying caregiver country of birth relative to Spain, and a household socio-economic background index. In addition, all school-level and socio-economic baseline variables that exhibited imbalance across treatment arms are included as controls. These covariates are included to improve precision and to address residual imbalance, consistent with the pre-analysis plan. Baseline outcomes are included whenever available.

Given the large number of related outcomes, statistical inference accounts for multiple hypothesis testing by reporting sharpened false discovery rate (FDR) q-values within pre-specified outcome families (Anderson, 2008). All main results reported in the paper correspond to this baseline specification.

E Additional Results on Socio-Emotional Skill Formation and Instructional Practices

Table E1: Heterogeneity of Impacts on Behavior-Based Socio-Emotional Skills Assessed by External Observers, by Socio-Economic Background

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Autonomy		Cooperation		Emotional Management		Responsibility		Thinking Abilities	
	Activity	Class	Activity	Class	Activity	Class	Activity	Class	Activity	Class
	Panel A. Students with lower socio-economic background									
Treated	0.102*	0.264***	0.070	0.174	0.053	0.208	0.142*	0.397***	-0.064	0.515***
	(0.058)	(0.099)	(0.063)	(0.112)	(0.059)	(0.128)	(0.075)	(0.085)	(0.056)	(0.175)
Control mean	-0.08	-0.10	0.00	-0.06	0.07	-0.13	-0.07	-0.12	0.03	-0.10
Observations	703	353	704	310	700	250	702	416	699	224
MHT q-val.	0.13	0.03	0.20	0.14	0.21	0.14	0.13	0.00	0.20	0.02
	Panel B. Students with higher socio-economic background									
Treated	-0.023	-0.176	0.065	-0.087	0.137**	0.067	-0.001	-0.124	-0.039	0.051
	(0.048)	(0.126)	(0.062)	(0.132)	(0.064)	(0.144)	(0.064)	(0.097)	(0.070)	(0.163)
Control mean	0.06	0.06	-0.00	0.04	-0.06	0.09	0.05	0.08	-0.03	0.07
Observations	938	469	940	426	931	329	930	549	934	300
MHT q-val.	1.00	1.00	1.00	1.00	0.56	1.00	1.00	1.00	1.00	1.00
Treated × Low SES	0.125	0.440***	0.005	0.261*	-0.085	0.141	0.142	0.522***	-0.025	0.464**
	(0.080)	(0.163)	(0.079)	(0.135)	(0.081)	(0.178)	(0.091)	(0.113)	(0.084)	(0.222)
Control mean	0.00	-0.00	-0.00	-0.00	0.00	0.00	-0.00	-0.00	-0.00	0.00
Observations	1641	822	1644	736	1631	579	1632	965	1633	524
Baseline outcome	No	No	No	No	No	No	No	No	No	No
Socio-demographic controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Unbalanced controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table reports heterogeneity in intent-to-treat estimates by socio-economic background of the effect of assignment to the intervention on behavior-based socio-emotional skill indices measured by external observers. Outcomes are standardized using the control group mean and standard deviation. Panel A reports the impacts on a split sample of students with below-the-median socio-economic background index, while Panel B reports the corresponding impacts of the above-the-median split sample. Last row reports the interacted heterogeneity coefficient on a full sample regression. Each column corresponds to a distinct skill domain measured either during a standardized group activity administered at endline or through in-classroom observations conducted during regular instruction. All specifications include school fixed effects and fixed effects for the randomization strata defined by school disadvantage level. Week-of-measurement fixed effects are included. Regressions control for pre-specified student socio-demographic characteristics and baseline covariates that exhibited imbalance across treatment arms. Baseline measures of the outcome are not available for these indices. Robust standard errors are reported in parentheses and are clustered at the school-grade level. The row labeled "MHT q-val." reports sharpened false discovery rate q-values accounting for multiple hypothesis testing within each outcome domain. Statistical significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table E2: Heterogeneity of Impacts on Behavior-Based Socio-Emotional Skills Assessed by External Observers, by Socio-Economic Background Discrete Index

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Autonomy		Cooperation		Emotional Management		Responsibility		Thinking Abilities	
	Activity	Class	Activity	Class	Activity	Class	Activity	Class	Activity	Class
	Panel A. Students with socio-economic background index ≤ 6									
Treated	0.102*	0.264***	0.070	0.174	0.053	0.208	0.142*	0.397***	-0.064	0.515***
	(0.058)	(0.099)	(0.063)	(0.112)	(0.059)	(0.128)	(0.075)	(0.085)	(0.056)	(0.175)
Control mean	-0.08	-0.10	0.00	-0.06	0.07	-0.13	-0.07	-0.12	0.03	-0.10
Observations	703	353	704	310	700	250	702	416	699	224
	Panel B. Students with socio-economic background index $= 7$									
Treated	0.026	0.001	0.078	-0.011	0.117	0.230	-0.004	-0.135	-0.037	0.155
	(0.060)	(0.178)	(0.065)	(0.175)	(0.076)	(0.175)	(0.078)	(0.115)	(0.069)	(0.201)
Control mean	0.02	-0.01	-0.01	0.07	-0.02	0.10	0.06	0.11	0.00	0.02
Observations	699	338	701	309	696	244	694	402	698	226
	Panel C. Students with socio-economic background index $= 8$									
Treated	-0.274**	-0.672**	-0.126	-0.137	0.151	-0.519	-0.008	-0.346	-0.080	-0.678
	(0.114)	(0.254)	(0.135)	(0.445)	(0.137)	(0.969)	(0.094)	(0.277)	(0.132)	(1.510)
Control mean	0.15	0.24	0.02	-0.03	-0.17	0.06	0.02	-0.00	-0.10	0.17
Observations	239	131	239	117	235	85	236	147	236	74

Notes: This table reports heterogeneity in intent-to-treat estimates by socio-economic background of the effect of assignment to the intervention on behavior-based socio-emotional skill indices measured by external observers. Outcomes are standardized using the control group mean and standard deviation. Panel A reports the impacts on a split sample of students with socio-economic index less or equal to 6, and Panel B and Panel C reports the corresponding impacts for split samples of students with socio-economic indices of 7 and 8, respectively. Last row reports the interacted heterogeneity coefficient on a full sample regression. Each column corresponds to a distinct skill domain measured either during a standardized group activity administered at endline or through in-classroom observations conducted during regular instruction. All specifications include school fixed effects and fixed effects for the randomization strata defined by school disadvantage level. Week-of-measurement fixed effects are included. Regressions control for pre-specified student socio-demographic characteristics and baseline covariates that exhibited imbalance across treatment arms. Baseline measures of the outcome are not available for these indices. Robust standard errors are reported in parentheses and are clustered at the school-grade level. The row labeled "MHT q-val." reports sharpened false discovery rate q-values accounting for multiple hypothesis testing within each outcome domain. Statistical significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table E3: Heterogeneity of Impacts on Behavior-Based Socio-Emotional Skills Assessed by External Observers, by baseline social engagement score

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Autonomy		Cooperation		Emotional Management		Responsibility		Thinking Abilities	
	Activity	Class	Activity	Class	Activity	Class	Activity	Class	Activity	Class
	Panel A. Students with low baseline social engagement									
Treated	0.130**	0.303**	0.110*	0.180	0.078	0.361**	0.177***	0.344***	0.103	0.223
	(0.053)	(0.125)	(0.056)	(0.114)	(0.077)	(0.160)	(0.059)	(0.099)	(0.068)	(0.168)
Control mean	-0.14	-0.20	-0.02	-0.09	0.09	-0.16	-0.04	-0.10	-0.02	-0.10
Observations	744	374	745	339	739	262	736	442	736	229
MHT q-val.	0.04	0.04	0.06	0.09	0.14	0.05	0.02	0.01	0.09	0.10
	Panel B. Students with high baseline social engagement									
Treated	0.002	-0.157	0.045	-0.110	0.118**	0.082	0.038	-0.070	-0.105	0.162
	(0.051)	(0.099)	(0.076)	(0.129)	(0.055)	(0.127)	(0.072)	(0.090)	(0.065)	(0.133)
Control mean	0.17	0.19	0.04	0.10	-0.06	0.11	0.08	0.10	0.10	0.12
Observations	897	448	899	397	892	317	896	523	897	295
MHT q-val.	1.00	0.54	0.86	0.86	0.52	0.86	0.86	0.86	0.54	0.66
Treated × Low social engag.	0.128*	0.460***	0.065	0.290**	-0.040	0.278	0.139*	0.414***	0.208**	0.062
	(0.076)	(0.144)	(0.089)	(0.146)	(0.085)	(0.189)	(0.082)	(0.112)	(0.089)	(0.225)
Control mean	0.03	0.02	0.02	0.01	0.01	-0.00	0.03	0.01	0.05	0.03
Observations	1641	822	1644	736	1631	579	1632	965	1633	524
Baseline outcome	No	No	No	No	No	No	No	No	No	No
Socio-demographic controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Unbalanced controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table reports heterogeneity in intent-to-treat estimates by baseline social engagement, as defined as the above/below median BESSI-45 score on the baseline social engagement domains. Outcomes are standardized using the control group mean and standard deviation. Panel A reports the impacts on a split sample of students with below the median baseline social engagement score, while Panel B reports the corresponding impacts for above the median students. Last row reports the interacted heterogeneity coefficient on a full sample regression. Each column corresponds to a distinct skill domain measured either during a standardized group activity administered at endline or through in-classroom observations conducted during regular instruction. All specifications include school fixed effects and fixed effects for the randomization strata defined by school disadvantage level. Week-of-measurement fixed effects are included. Regressions control for pre-specified student socio-demographic characteristics and baseline covariates that exhibited imbalance across treatment arms. Baseline measures of the outcome are not available for these indices. Robust standard errors are reported in parentheses and are clustered at the school-grade level. The row labeled "MHT q-val." reports sharpened false discovery rate q-values accounting for multiple hypothesis testing within each outcome domain. Statistical significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table E4: Heterogeneity of Impacts on Behavior-Based Socio-Emotional Skills Assessed by External Observers, by baseline emotional management score

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Autonomy		Cooperation		Emotional Management		Responsibility		Thinking Abilities	
	Activity	Class	Activity	Class	Activity	Class	Activity	Class	Activity	Class
	Panel A. Students with low baseline emotional management									
Treated	0.073	0.269***	0.019	0.172	0.022	0.172	0.085	0.339***	-0.022	0.456***
	(0.046)	(0.100)	(0.050)	(0.116)	(0.060)	(0.149)	(0.066)	(0.087)	(0.063)	(0.151)
Control mean	0.04	-0.04	0.08	-0.08	0.00	-0.09	0.02	-0.13	0.03	-0.07
Observations	851	409	852	375	844	294	846	490	841	257
MHT q-val.	0.25	0.03	0.49	0.25	0.49	0.34	0.30	0.00	0.49	0.02
	Panel B. Students with high emotional management									
Treated	0.029	-0.111	0.113	-0.189	0.134**	0.137	0.024	-0.212**	-0.113*	0.117
	(0.057)	(0.134)	(0.075)	(0.133)	(0.062)	(0.143)	(0.068)	(0.098)	(0.057)	(0.138)
Control mean	0.02	0.07	-0.05	0.12	0.01	0.09	0.03	0.16	0.07	0.12
Observations	790	413	792	361	787	285	786	475	792	267
MHT q-val.	0.52	0.47	0.29	0.29	0.21	0.47	0.56	0.21	0.21	0.47
Treated × Low emotional manag.	0.044	0.380**	-0.094	0.361**	-0.112	0.035	0.061	0.552***	0.091	0.339*
	(0.076)	(0.146)	(0.082)	(0.157)	(0.073)	(0.189)	(0.088)	(0.119)	(0.071)	(0.176)
Control mean	0.03	0.02	0.02	0.01	0.01	-0.00	0.03	0.01	0.05	0.03
Observations	1641	822	1644	736	1631	579	1632	965	1633	524
Baseline outcome	No	No	No	No	No	No	No	No	No	No
Socio-demographic controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Unbalanced controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table reports heterogeneity in intent-to-treat estimates by baseline emotional management, as defined as the above/below median BESSI-45 score on the baseline emotional management domains. Outcomes are standardized using the control group mean and standard deviation. Panel A reports the impacts on a split sample of students with below the median baseline emotional management score, while Panel B reports the corresponding impacts for above the median students. Last row reports the interacted heterogeneity coefficient on a full sample regression. Each column corresponds to a distinct skill domain measured either during a standardized group activity administered at endline or through in-classroom observations conducted during regular instruction. All specifications include school fixed effects and fixed effects for the randomization strata defined by school disadvantage level. Week-of-measurement fixed effects are included. Regressions control for pre-specified student socio-demographic characteristics and baseline covariates that exhibited imbalance across treatment arms. Baseline measures of the outcome are not available for these indices. Robust standard errors are reported in parentheses and are clustered at the school-grade level. The row labeled "MHT q-val." reports sharpened false discovery rate q-values accounting for multiple hypothesis testing within each outcome domain. Statistical significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table E5: Heterogeneity in Instructional Practice Responses by Classroom Socio-Economic Composition

	(1)	(2)	(3)	(4)	(5)	(6)
	Instructional Organization		**Instructional interaction style**		**Regulation of learning and behavior**	
	Whole class instruction (%)	Focus on content (%)	Interactive instruct. style (%)	Motivational support (%)	Encourages student agency (%)	Corrects student behavior (%)
	Panel A. High share of disadvantaged students in the class					
Treated	4.414	-13.546***	15.979**	6.953	-8.395	-2.329
	(11.737)	(4.855)	(5.908)	(7.950)	(9.086)	(2.537)
Control mean	49.31	61.81	29.17	24.31	25.69	8.33
Observations	111	111	111	111	111	111
MHT q-val.	0.87	0.03	0.03	0.45	0.45	0.45
	Panel B. Low share of disadvantaged students in the class					
Treated	7.435	-0.424	0.396	18.147**	13.217**	-17.520*
	(11.329)	(11.523)	(6.039)	(8.170)	(6.150)	(10.194)
Control mean	29.82	53.57	28.57	21.96	16.25	11.25
Observations	113	113	113	113	113	113
MHT q-val.	0.63	0.94	0.94	0.13	0.13	0.14
Treated × High share of underpr. stud.	-3.020	-13.122	15.584*	-11.194	-21.612*	15.191
	(16.238)	(12.486)	(8.412)	(11.351)	(10.907)	(10.498)
Control mean	36.44	56.37	28.77	22.76	19.46	10.26
Observations	224	224	224	224	224	224
Baseline outcome	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table reports heterogeneity in intent-to-treat estimates of the effect of assignment to the intervention on observed instructional practices, by classroom socio-economic composition. Outcomes are defined as the share of five-minute observation intervals in which the corresponding instructional practice is recorded. Panel A reports estimates for classrooms with a high share of disadvantaged students, and Panel B reports estimates for classrooms with a low share of disadvantaged students. The final rows report interaction estimates from a pooled regression. All specifications include school fixed effects and week-of-measurement fixed effects are included. Regressions control for baseline values of the outcome when available. Robust standard errors are reported in parentheses and are clustered at the school-grade level. Statistical significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

F Additional Material

Barcelona, February 13 2023

To whom it may concern,

As a co-founder of the methodology being implemented in this project, I will recuse myself from actively participating in the data analysis to ensure impartiality and maintain the integrity of the results. My role in this project will be limited to providing guidance and support in the design and implementation of study**,** while leaving the data analysis to the designated team members.

Caterina Calsamiglia

