working paper

GEAR
GRADUATE
PROGRAM
IN APPLIED
ECONOMIC
RESEARCH

Gender Gaps in Socioemotional Skills: Evidence from the Classroom

Ece Yagman

Working paper:
2025-01

Universitat Autònoma
de Barcelona

Departament
d'Economia Aplicada

UAB

Avinguda de l'Eix Central
Edifici B2
Campus de la UAB
08193 Bellaterra
(Cerdanyola del Vallès)
Barcelona · Spain
Tel. +34 93 581 16 80
Fax +34 93 581 22 92
d.econ.aplicada@uab.cat
www.uab.cat/departament/
economia-aplicada/

Coordinator: Rosella Nicolini (rosella.nicolini@uab.cat)

Gender Gaps in Socioemotional Skills: Evidence from the Classroom

ECE YAGMAN
Universitat Autonoma de Barcelona

November 14, 2025

Abstract

Using repeated classroom assessments of socioemotional skills for more than 1,100 adolescents in 40 secondary schools, this paper documents significant gender differences in self- and peer evaluations. Conditional on teacher and peer ratings, female students systematically underrate themselves in socioemotional domains culturally stereotyped as masculine—specifically Emotional Management and Thinking Abilities. Conversely, peer evaluations reveal a robust and novel female–female premium: female students consistently assign significantly higher ratings (0.21–0.37 SD) to their female classmates across all socioemotional domains. These patterns replicate when teacher scores are replaced by external observers, in a lab-in-the-field setting, and under a validated survey-based measure of socioemotional skills. Whereas prior literature extensively documents gender gaps in self- and peer-assessments of cognitive skills, this study provides novel evidence of similar asymmetries in adolescents' assessments of socioemotional competencies. These findings expand the scope of gendered evaluation dynamics in formative educational environments and highlight early adolescence as a policy-relevant window for targeted interventions addressing gender gaps in self and peer perceptions.

1 Introduction

Over the past half-century, women have overtaken their male peers in educational outcomes, including grade point averages, graduation rates, and university enrollment. Yet these academic gains have not translated into corresponding advantages in the labor market: a substantial gender wage gap persists at entry and widens across the career lifecycle (Delaney and Devereux, 2021; Goldin et al., 2006; Goldin, 2014; Blau and Kahn, 2017). This enduring gap suggests that conventional human-capital indicators do not fully explain the whole story. Recent work in economics points to another set of mechanisms: how individuals value their own abilities, and how those abilities are perceived and evaluated by others, may shape trajectories at least as powerfully as formal qualifications. However, it remains unclear to what extent these perceptions—and their influence on life outcomes—are themselves shaped by gender, and at what point such dynamics start to matter.

A growing literature in economics emphasizes the rising importance of socioemotional—often termed "soft" or "non-cognitive"—skills. Attributes such as teamwork, adaptability, and emotional regulation predict educational attainment, earnings, and well-being, often on par with cognitive skills (Heckman and Kautz, 2012). As automation steadily erodes the market premium on routine technical tasks, these interpersonal and self-regulatory skills have become ever more valuable (Deming, 2017). Existing research documents gender differences in socioemotional skill levels. Girls, for instance, are often rated higher in cooperation and self-discipline (Duckworth and Seligman, 2006; Ogden et al., 2023). Yet, despite the clear economic significance of these skills and the documented gender differences in their levels, surprisingly little is known about whether gender biases influence their evaluation. In particular, whether gender dynamics shape the evaluation of adolescents' socioemotional abilities, either through self-assessment or in external assessments, remains unclear. This paper addresses precisely this question by examining gender differences in how adolescents' socioemotional skills are evaluated within educational settings.

Early adolescence offers a critical and revealing context for studying this issue, as it marks a period when neuroplasticity peaks and social-affective learning is particularly sensitive to external feedback (Steinberg, 2014; Yeager, 2017). Importantly, gender differences in key behavioral traits such as competitiveness emerge prominently around puberty and are shaped by socialization processes (Andersen et al., 2013). Therefore, classroom environments provide a natural setting for observing, fostering, and assessing these competencies.

However, while schools offer a promising environment for socioemotional learning, it is far from clear that the evaluation of these skills is neutral or unbiased. A substantial body of research has already shown that assessment itself is not gender-neutral. For instance, studies in higher education reveal that female instructors receive lower student evaluations than their male counterparts despite similar objective performance (Boring, 2017; Mengel et al., 2019). In experimental settings, male peer graders give lower scores to work attributed to female students (Saygin and Knight, 2023), and women themselves tend to under-promote their own performance (Exley and Kessler, 2022). These findings collectively illustrate a gendered structure of academic feedback. However, nearly all prior studies have focused exclusively on cognitive outcomes, such as grades or standardized test scores. Whether similar asymmetries extend to adolescents' socioemotional abilities—the very skills increasingly valued by employers—remains an open and policy-relevant question.

To address this question, we leverage a dataset following 1,102 students in 40 secondary schools in Catalonia, Spain, capturing repeated evaluations of socioemotional skills across two academic terms. The dataset includes assessments from four distinct rater groups: students themselves, their classmates, teachers, and external observers trained for impartiality. Each assessment covers five socioemotional domains—Responsibility, Autonomy, Cooperation, Emotional Management, and Thinking Abilities—anchored to directly observable behaviors. This multi-rater framework allows us to compare self- and external perceptions, disentangle peer dynamics from teacher evaluations, and systematically examine gender-based differences in skill assessments. By triangulating across these multiple perspectives, we provide a picture of gendered dynamics in socioemotional skill assessment. To our knowledge, no prior study has combined such a comprehensive array of behavioral evaluations within the same adolescent population, making our contribution novel to the literature.

We complement the classroom analysis with two additional components: a lab-in-the-field exercise and a validated survey inventory to measure socioemotional skills. The lab-in-the-field activity was implemented across all participating schools towards the end of the academic year. In this standardized exercise, students were randomly assigned to groups to complete a collaborative and competitive task. Immediately after the activity, both students and trained external observers evaluated each participant's performance using a concise 10-item questionnaire. This instrument was specifically designed to capture the same five socioemotional domains assessed in the classroom, allowing for a direct comparison between the classroom-based behavioral ratings and the survey-based assessments in a different context.

At endline, the Behavioral, Emotional, and Social Skills Inventory (BESSI) was administered to obtain 360-degree measures of students' socioemotional skills (Soto et al., 2022). Students completed a 20-item self-report and provided 20-item observer-reports for two classmates; teachers also submitted observer-reports for a subsample of students. The BESSI assesses five domains of socioemotional skills, conceptualized as functional capacities—"how well" behaviors are enacted when required. Collected independently of classroom activities, the instrument has been validated for adolescent self- and observer-reports. Including BESSI thus allows the study to triangulate findings across measurement methods and rater types within a five-factor framework.[1]

We report three central findings. First, conditional on evaluations by teachers and peers, female students consistently underrate their own skills in domains culturally coded as masculine—namely, *Emotional Management* and *Thinking Abilities*. This self-critical pattern aligns with existing evidence that women systematically underrate their performance on tasks stereotypically associated with males, beginning as early as middle school (Exley and Kessler, 2022). Our results suggest that these gender gaps in self-evaluations begin to emerge around puberty, echoing the developmental

[1] A multitude of socioemotional skills taxonomies exist (see Explore SEL database). Despite differences in labels and definitions, the consensus is that these skills can systematically be mapped onto five broad and non-overlapping domains, analogous to the Big Five (Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness), to explore how personality affects behavior and shapes life outcomes (e.g., Abrahams et al. (2019); OECD (2015); John et al. (2008)).

timing found by Andersen et al. (2013).

Second, we uncover a notable asymmetry in peer evaluations: female students systematically rate their female classmates significantly higher across all socioemotional domains, generating a robust female-female premium of between 0.21 and 0.37 standard deviations. In contrast, male evaluators show no systematic gender difference. This asymmetry suggests that, while girls are relatively more critical of their own abilities, they do not project this self-criticism onto their female peers. Instead, conditional on both evaluator's and evaluee's self and teacher scores, female peers extend an in-group premium to their female classmates. This finding adds fresh evidence by showing that evaluation dynamics among adolescent girls can operate through positive peer recognition, not solely through self-deprecation.

Third, these gendered patterns in peer assessments persist when teacher ratings are replaced by external observers—a more neutral benchmark—and they are replicated both in the standardized lab-in-the-field exercise and in the endline survey. In the lab, a concise 10-item instrument reproduces the classroom results. At endline, the BESSI-20 self and observer reports collected independently of classroom activities replicates a similar pattern. Strikingly, the gender gaps documented over six months of classroom-based behavioral ratings reappear in both the brief lab and endline survey, indicating that these tendencies can be captured by extended behavioral assessments and conventional survey measures alike.

Taken together, our findings demonstrate that gender gaps in socioemotional skill evaluations are already present early in adolescence and persist across different evaluation contexts. The gendered nature of these assessments may ultimately feed into the enduring disparities observed in high-stakes academic and labor market settings. Indeed, research has shown that students' over- or under-estimation of their own abilities accounts for a sizable share of gender gaps in college major choice and expected earnings (Reuben et al., 2017). Recent work by Benson et al. (2024) shows that, in a large North American retail chain, women receive substantially lower subjective "potential" ratings than men, despite stronger job performance, with these assessments accounting for roughly half of the gender promotion gap. Changing who evaluates is not, by itself, a remedy:

exploiting randomized variation in the gender composition of academic committees in Spain and Italy, Bagues et al. (2017) find that greater female representation does not systematically benefit female candidates and may even backfire under certain group dynamics.

By comparison, our study uncovers a notable *female–female* premium in peer assessments during adolescence: female students systematically assign higher scores to their female classmates, even as they remain more self-critical. While the finding that girls are tough on themselves aligns with broader literature, the evidence that they consistently value their female classmates more highly represents a novel contribution. This insight opens promising avenues for future research and policy interventions: explicitly informing female students about these asymmetries, highlighting the "blind-spot" competencies recognized by their female peers, may help mitigate overly critical self-assessments and strengthen self-perceptions. By documenting these early, context-specific patterns, we provide empirical evidence on an underexplored channel through which gendered evaluation dynamics emerge and potentially shift over time, sometimes reinforcing, yet in other instances diverging from, the biases observed in adult, high-stakes environments.

The remainder of the paper proceeds as follows. Section 2 reviews related literature. Section 3 describes the study context. Section 4 presents descriptive statistics, empirical strategies, and main findings. Section 5 provides robustness checks. Section 6 explores heterogeneity by age group, and Section 7 concludes with policy implications and avenues for future research.

2 Related Literature

A substantial body of literature documents persistent gender disparities in labor market outcomes: women, on average, earn less than men, climb the promotion ladder more slowly, and remain under-represented in senior positions (Goldin, 2014; Marianne, 2011; Blau and Kahn, 2017). A complementary experimental and field literature attributes part of this gap to systematic gender differences in self-beliefs, preferences, and behaviors. Women tend to exhibit lower confidence in their abilities relative to equally able men (Beyer and Bowden, 1997; Barber and Odean, 2001; Soll and Klayman, 2004) and rate their own performance more critically, even in the absence of incentives to self-promote (Exley and Kessler, 2022). Laboratory experiments reveal that while men significantly increase effort as competitive pressure intensifies, women's performance remains relatively unchanged (Gneezy et al., 2003; Niederle and Vesterlund, 2011). Furthermore, women are also less willing to enter mixed-gender tournaments on stereotypically male tasks (Niederle and Vesterlund, 2007; Exley and Kessler, 2022; Buser et al., 2023). Field and experimental evidence also shows that women perform relatively better when stakes are lower (Azmat et al., 2016), are less likely to speak up or contribute ideas in group discussions (Coffman, 2014), and systematically negotiate or ask for raises and promotions less frequently than their male counterparts (Babcock and Laschever, 2021; Hernandez-Arenaz and Iriberri, 2019; Recalde and Vesterlund, 2023; Roussille, 2024). Importantly, much of what is perceived as a gender gap in social preferences is rooted in widely shared beliefs rather than actual behavioral differences: both men and women expect women to be more generous and equality-oriented than men, yet experimental evidence finds these behavioral gaps to be negligible (Exley et al., 2024). Collectively, these findings suggest that gendered differences in self-beliefs and preferences all contribute to persistent disparities in the labor market.

If self-beliefs matter, the natural next question is how they are formed. A large body of work points to the formative role of feedback and assessment, ranging from individuals' own self-perception to evaluations by others. Evidence across disciplines shows that women internalize critical self-assessments more from an early age. In a cohort of UK biology undergraduates, Langan et al. (2008) report that women assign themselves lower marks than achievement-matched men; Rust et al. (2003) describe a similar pattern in psychology. Qualitative work among Spanish engineering students confirms that women perceive higher performance thresholds as the minimum requirement for success (Torres-Guijarro and Bengoechea, 2017). Beyond academia, women are less likely to claim proficiency in programming languages on CVs (Murciano-Goroff, 2022). Exley and Kessler (2022) extend the evidence to adolescents, showing that gender gaps in self-evaluation on a male-typed task emerge by age eleven and remain stable through high school.

Recent evidence on socioemotional skills reinforces these patterns. Using harmonized data on

42,000 adolescents and young adults in 17 African countries, Ajayi et al. (2022) find a male premium of 0.15 SD in self-reported SES, concentrated in agentic traits such as emotional regulation and problem solving & decision making, with no compensating female advantage in communal skills.[2] Whether this gap reflects true skill differences or biased self-assessment is addressed by Cassidy et al. (2024), who combine self-reports and behavioral measures of 14 distinct socioemotional constructs in urban Tanzania. The self-report data replicate a sizable male advantage, yet this gap vanishes almost entirely once skills are measured through behaviors: men overstate their competence, whereas women's survey answers track their behavior far more closely. Such mismeasurement can have real consequences. Persistent gender differences in self-beliefs and aspirations help explain why, even as women surpass men in educational attainment, they remain underrepresented in high-earning fields and continue to face wage gaps in the labor market (Delaney and Devereux, 2021).

Turning from internal to external evaluations, peer assessment is widely promoted as a pedagogical tool that fosters accountability and team skills, yet its susceptibility to gender bias remains contested. Early syntheses concluded that peer marks closely track expert judgments and show little systematic bias (Falchikov and Magin, 1997). Tucker (2014) similarly detects no systematic bias but notes marginally higher ratings for women. In a large-scale study of oral presentations, however, Langan et al. (2005) find that male students award slightly higher scores to male presenters, while female assessors appear gender-neutral. More recent work reaches divergent conclusions: analyzing 1,650 peer ratings in economics courses, Espey (2022) attributes higher average scores for women to better observable contributions rather than rater bias, whereas Smith and Wooten (2024) apply a censored-rubric estimator to 2,000 ratings and uncover a significant negative bias against female students by both male and female peers. Saygin and Knight (2023) exploit random assignment of anonymous peer graders and show that male graders assign lower scores to submissions with female-sounding names, while female graders are gender-neutral. The heterogeneity across studies suggests that peer bias is context-dependent, varying with task type, grading

[2] See Eagly and Wood (2012) for a discussion of agentic skills that are culturally associated with men, and communal skills (e.g., cooperation, empathy) with women.

rubric, and the gender composition of teams. Because peer reviews can also play a role later in life (e.g., workplace performance appraisals, referral networks, etc.), understanding these dynamics in educational settings remains important.

Recent evidence also documents bias in student evaluations of teaching (SET) in higher education, where such evaluations can play a pivotal role in hiring, promotion, and tenure decisions. Randomized evidence shows that perceived instructor gender alone can sway evaluations: in an online course MacNell et al. (2015) assigned identical teaching to avatars with male- or female-sounding names, and students rated the "female instructor" lower on every dimension, including objectively verifiable ones such as turnaround time. Using a large dataset of student evaluations from a natural experiment at a French university and a randomized experiment at a U.S. university, Boring and Ottoboni (2016) show that instructors believed to be male receive significantly higher teaching evaluations than their female counterparts, with male students driving the bias in France and female students in the U.S. Similar gaps emerge in Australian panel data (Fan et al., 2019), German business schools (Wagner et al., 2016), and a multi-institution study with within-course fixed effects (Mengel et al., 2019). Notably, Fan et al. (2019) find that the bias attenuates in departments with a higher share of women, indicating that representation can dampen—but not eliminate—discriminatory evaluations. Collectively, these studies document a systematic penalty for women that operates even when course content, grades, and instructor quality are held constant, raising serious concerns about the reliability of SET as a decision-making tool for career decisions.

Bias can also flow the opposite way, from teachers to pupils. Using Israeli high-school data, Lavy (2008) reports that teachers favor girls in languages and boys in mathematics, with the sign of the gap depending on teacher gender. Cornwell et al. (2013) exploit U.S. primary-school data and show that girls outperform boys on teacher-graded work relative to standardized tests, implying teacher bias or differential skill measurement. By contrast, Hinnerich et al. (2011) find no systematic gender discrimination in Swedish high schools once identical scripts are double-marked, and Avitzour et al. (2020) report null effects in Israeli primary schools when each exam is presented with both male and female identifiers. Despite the heterogeneity, the balance of evidence indicates that teacher judgments can amplify or dampen gender gaps and that the direction of bias may vary by subject, teacher gender, and grading format.

Existing literature, thus, establishes that gender-linked differences in psychological traits and evaluative bias jointly distort how ability is judged. Across different settings, the evidence points to two recurring facts: (i) women underrate their own ability, particularly on stereotypically-male or competitive tasks, and (ii) external evaluators, whether classmates or students, often hold women to stricter standards or discount female achievements altogether. These asymmetries surface as early as primary school, persist through university, and extend from cognitive outcomes to self-reported socioemotional skills.

Building on this foundation, our contribution is threefold. First, we observe the same group of students drawn from four academic cohorts (1st to 4th ESO), tracking the same underlying construct (five socioemotional competencies) through four independent lenses: students' self-ratings, peer ratings, teacher ratings, and assessments by trained external observers. By anchoring all raters to a shared behavioral rubric and collecting repeated assessments throughout the school year, we place each evaluator's scores on a common scale, allowing us to map systematic discrepancies across evaluators.

Second, whereas most prior work explores gender gaps in cognitive performance assessments, we turn the spotlight to socioemotional skills. Early adolescence is precisely when these competencies crystallize and when feedback is most likely to mold self-beliefs; yet we know little about whether the gender biases documented for test scores and course evaluations spill over to the soft skills that the labor market increasingly rewards.

Third, we reveal a notable asymmetry in female assessments: while females are systematically harsh on themselves in the two domains most culturally coded as male, *Emotion Management* and *Thinking Abilities*, they do not project this discount onto their female classmates. Instead, they award significantly higher ratings to other females across *all* socioemotional skills. This combination of self-undervaluation and in-group endorsement provides new evidence on how gendered perceptions are both internalized and expressed in everyday classroom interactions.

3 Study Context

3.1 The Program

We use a novel classroom-level dataset from the 2022–2023 academic year. In 88 classrooms across 40 public secondary schools in Catalonia, Spain, teachers integrated *Pentabilities*, a formative assessment program for socioemotional learning, into regular instruction.[3] Before the school year, schools were invited to adopt the program; participating schools opted in voluntarily. Teachers completed a blended training package consisting of four hours of asynchronous online modules followed by a four-hour in-person workshop led by experienced mentors. From January to June 2023, teachers received monthly mentoring while implementing the program and they also participated in two focus groups. [4]

As Figure 1 illustrates, *Pentabilities* conceptualizes socioemotional learning as a set of 35 discrete, observable behaviors grouped into five domains: *Responsibility*, *Autonomy and Initiative*, *Cooperation*, *Emotional Management*, and *Thinking Abilities*. Each behavior is rated on a 5-point scale (1: not very well, 5: extremely well) and from a 360-degree perspective that combines teacher, peer, and self-ratings. Teachers introduce the framework within their subject areas (e.g., mathematics, science, languages). Some classrooms were led by a single participating teacher, others by several.

[3] Developed by veteran teachers and education specialists, the program has been implemented for more than a decade from primary through master's level in Spain and abroad; see Pentabilities.

[4] Further details on program design and the accompanying impact evaluation are in the working paper *The Impact of Formative Assessment of Behavior-Based Socioemotional Skills on Students' Outcomes in the Short and Long Run*.

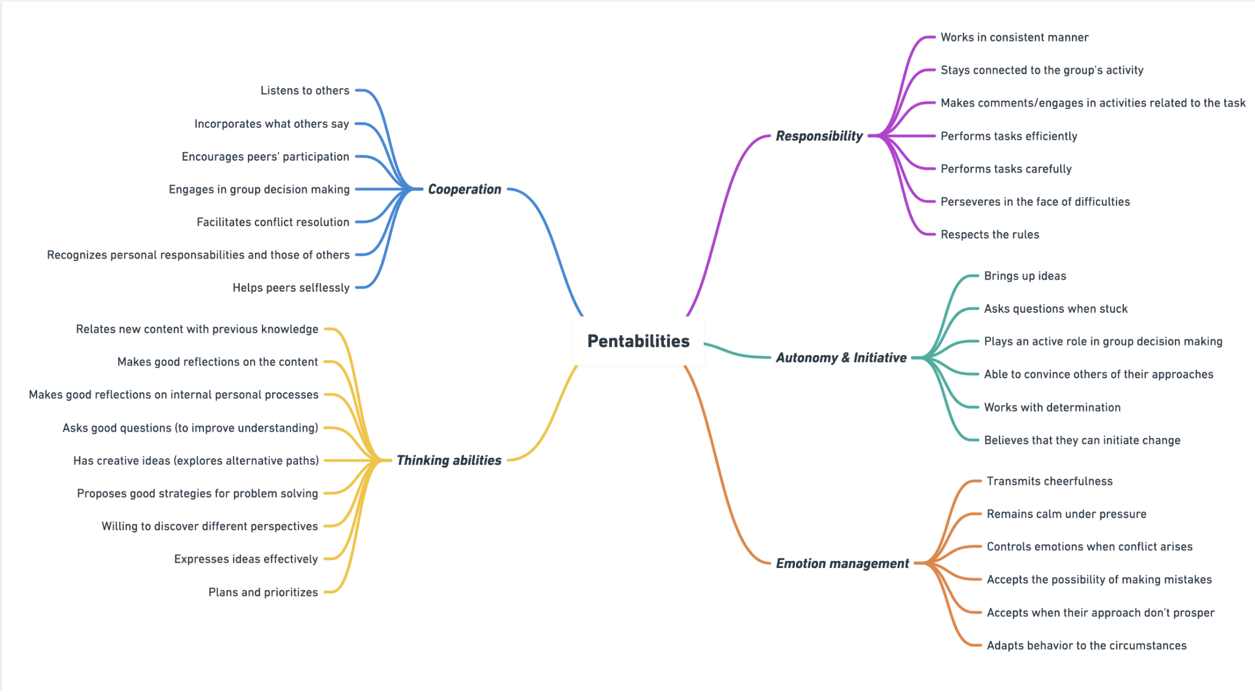

Pentabilities
Cooperation
Listens to others
Incorporates what others say
Encourages peers' participation
Engages in group decision making
Facilitates conflict resolution
Recognizes personal responsabilities and those of others
Helps peers selflessly
Thinking abilities
Relates new content with previous knowledge
Makes good reflections on the content
Makes good reflections on internal personal processes
Asks good questions (to improve understanding)
Has creative ideas (explores alternative paths)
Proposes good strategies for problem solving
Willing to discover different perspectives
Expresses ideas effectively
Plans and prioritizes
Responsibility
Works in consistent manner
Stays connected to the group's activity
Makes comments/engages in activities related to the task
Performs tasks efficiently
Performs tasks carefully
Perseveres in the face of difficulties
Respects the rules
Autonomy & Initiative
Brings up ideas
Asks questions when stuck
Plays an active role in group decision making
Able to convince others of their approaches
Works with determination
Believes that they can initiate change
Emotion management
Transmits cheerfulness
Remains calm under pressure
Controls emotions when conflict arises
Accepts the possibility of making mistakes
Accepts when their approach don't prosper
Adapts behavior to the circumstances

Figure 1: Map of Pentabilities-35 behaviors and 5 domains

During each class, teachers engage the students in collaborative and task-based activities designed to elicit the target behaviors. Students are randomly assigned to small groups of three to five; depending on the task, these groups either stay together for several sessions or are reshuffled, a strategy that limits rating bias arising from close friendships. Using the *Pentabilities* web or mobile app, teachers record their ratings and also ask the students to submit self- and peer assessments at the end of each lesson. Peer ratings are within-group: each student rates all other members of their group, not the entire class. Because groups are reshuffled across sessions, each student both rates and is rated by multiple, changing classmates over time.

All ratings are strictly confidential—never disclosed to classmates, have no bearing on formal grades, and displayed only in anonymized dashboards. This privacy framework encourages honest feedback that supports socioemotional development: students cannot trace a score back to a particular peer and, in classes with multiple instructors, cannot discern which teacher contributed which rating.[5]

[5] In multi-teacher classrooms students view only the average teacher score.

The ultimate purpose of evaluating these behaviors is to collate them in feedback reports. The digital platform aggregates data into individual dashboards that display each student's trajectory by rater type. On average, once per term teachers hold dedicated feedback sessions–either one-to-one or in small groups–in which students interpret their reports, reflect on discrepancies across raters, and set goals for the next cycle.

To benchmark these in-class assessments, two external observers (EOs), recruited and managed by an independent organization, visited each classroom.[6] In every visit they randomly selected 8–10 students (split evenly between observers with a small overlap for inter-rater reliability). Using the same 35-behavior rubric, EOs rated only those behaviors actually displayed during that lesson; their scores were stored solely for research purposes and never shared with teachers or students, and excluded from feedback reports.[7]

Taken together, the multi-rater design yields a unique, classroom-level dataset of socioemotional indicators observed across subjects and activities. While teacher ratings could inevitably reflect subjective judgment, parallel scores from independent external observers offer a more objective reference point, allowing us to run the analysis with an alternative benchmark.

3.2 Standardized Activity

To test the validity of our classroom findings, we draw on data from a one-hour "lab-in-the-field" exercise conducted in every participating school towards the end of the academic year, between May and June 2023. Designed to be context-independent, this session was held outside regular lesson times and placed all students in a tightly scripted collaboration and competition task, with EOs rating socioemotional skills in place of teachers. The resulting dataset provides a benchmark against which to assess the robustness of our main results.

The activity unfolded in three phases. First, after receiving standardized instructions, students were randomly assigned to groups of three to five and given ten minutes to construct the tallest

[6] The external organization in charge of hiring and training was Empieza por Educar, the Spanish version of the organization Teach for All.

[7] While teachers aim to evaluate every student, EOs rate at most ten students per classroom. Because of this, analysis using the EO scores gives us a subset of the sample.

free-standing tower capable of suspending marshmallows at least 20 cm above the desk, using a fixed kit of materials (straws, sticks, elastic bands, interlocking bricks).[8] In the second phase, teams completed a two-stage reflection cycle: one minute of individual self-assessment followed by a five-minute group discussion on strategy. In the final phase, groups had an additional five minutes to improve their structures. Scores were then calculated and the highest-scoring group was recognized as the winner. The activity was not incentivized beyond this recognition and did not affect students' official grades.

For reasons of brevity and practicality, the standardized activity adopted a different measurement tool for socioemotional skills: a concise 10-item questionnaire covering all five socioemotional domains. Upon completing the activity, both EOs and students privately rated how well each participant (or their peers) demonstrated the relevant behaviors on a 5-point scale.[9] Students completed the survey for themselves and for every peer in their group, while external observers rated the students they observed. Because this post-exercise survey produced self, peer, and external ratings using identical wording, it serves as the 360-degree measure in our robustness analysis.

3.3 Data Description

We use student-level data from 40 secondary schools and 88 classrooms. While the dataset contains multiple observations per student, reflecting evaluations captured at several points during the year, the present analysis aggregates these repeated measures into a single cross-sectional snapshot for each student.[10] Table 1 summarizes the main descriptive statistics for the full sample and separately by student gender.

The sample includes 1,102 students (515 girls and 587 boys), with an average age of 14. Most students were born in Spain (85% overall), and caregiver backgrounds look similar across groups: about 60% of primary and secondary caregivers were born in Spain, and roughly 40% of primary caregivers report a university degree.

[8] See Figures D.1 and D.2 in Appendix D for examples. Each marshmallow successfully suspended earned points; each additional piece of material used deducted points.

[9] Details on questionnaire construction and measurement are provided in Appendix A.2.

[10] This approach facilitates clear comparisons across individuals but does not yet exploit the full panel dimension of the data, which is reserved for future work.

Turning to school characteristics, more than 85% of them are located in the province of Barcelona. The study is predominantly situated in high-complexity institutions, which serve higher-need populations with larger shares of families with low parental education, immigrant status, or minimum-income receipt and are considered academically and socially at risk. Most schools are classified as either "High" or "Medium-High" on the official complexity index.[11] In Catalonia, compulsory secondary education is organized into two cycles over four academic years: the first cycle comprises 1st and 2nd years of ESO (ages 12–14), while the second cycle covers 3rd and 4th years (ages 14–16). About 70% of our sample is enrolled in the first cycle, with the remainder in the second cycle. Finally, on average, 66% of students in a classroom consented to participate. Among the consented group, the within-class gender split is around 48% girls and 52% boys.[12]

Teacher and external observer (EO) characteristics are presented in Table 2. The teaching staff is predominantly female, with women outnumbering men two to one, and an average age of around 42. There are nine external observers in total, the majority of whom are female.

[11] The complexity index ranges from 1 ("Low-Mid Complexity") to 4 ("Very High Complexity"); see Moreno and Iñesta (2021) for details.

[12] Personal characteristics, such as gender and age, are available only for students whose parents provided consent and who completed either the baseline or endline surveys. The Pentabilities protocol was implemented classroom-wide, but non-consented students are excluded from the analysis dataset. Reported shares thus describe the consented subsample rather than the true classroom composition.

Table 1: Student, Caregiver and School Characteristics

	All	Female	Male	Mean Difference
Panel A: Student Characteristics				
Student Age	13.89	13.87	13.91	-0.04
	(1.04)	(1.02)	(1.06)	(0.06)
Born in Spain	0.85	0.84	0.86	-0.02
	(0.35)	(0.37)	(0.34)	(0.02)
Panel B: Primary Caregiver Characteristics				
Born in Spain	0.60	0.58	0.61	-0.03
	(0.49)	(0.49)	(0.49)	(0.03)
Attended University	0.40	0.38	0.41	-0.04
	(0.49)	(0.48)	(0.49)	(0.03)
I don’t know (if attended university)	0.25	0.23	0.28	-0.05**
	(0.43)	(0.42)	(0.45)	(0.02)
Panel C: Secondary Caregiver Characteristics				
Born in Spain	0.62	0.60	0.63	-0.04
	(0.49)	(0.49)	(0.48)	(0.03)
Attended University	0.37	0.42	0.32	0.10***
	(0.48)	(0.49)	(0.47)	(0.03)
I don’t know (if attended university)	0.30	0.28	0.31	-0.03
	(0.46)	(0.45)	(0.46)	(0.03)
Panel D: School Characteristics				
Barcelona	0.86	0.84	0.87	-0.02
	(0.35)	(0.36)	(0.34)	(0.02)
Girona	0.03	0.03	0.03	0.00
	(0.16)	(0.16)	(0.16)	(0.01)
Lleida	0.09	0.10	0.08	0.02
	(0.29)	(0.30)	(0.27)	(0.01)
Tarragona	0.03	0.03	0.03	0.00
	(0.16)	(0.16)	(0.16)	(0.00)
Complexity: Medium-Low	0.32	0.29	0.35	-0.06**
	(0.47)	(0.45)	(0.48)	(0.03)
Complexity: High	0.33	0.36	0.31	0.05**
	(0.47)	(0.48)	(0.46)	(0.02)
Complexity: Medium-High	0.26	0.25	0.27	-0.01
	(0.44)	(0.43)	(0.44)	(0.03)
Complexity: Very High	0.09	0.10	0.07	0.02
	(0.28)	(0.30)	(0.26)	(0.02)
1st ESO	0.43	0.43	0.43	0.00
	(0.50)	(0.50)	(0.50)	(0.03)
2nd ESO	0.25	0.24	0.26	-0.02
	(0.43)	(0.43)	(0.44)	(0.02)
3rd ESO	0.27	0.28	0.25	0.03
	(0.44)	(0.45)	(0.44)	(0.03)
4th ESO	0.05	0.05	0.05	-0.00
	(0.22)	(0.22)	(0.23)	(0.01)
Panel E: Classroom Consent Rate by Gender				
Classroom consent rate	66.36	0.48	0.52	-0.04
	(21.39)	(0.15)	(0.15)	(0.04)
Observations	1102	515	587	

Notes: Student and caregiver information was collected at the baseline and endline surveys. School information comes from administrative data. Differences estimated with classroom-level clustered SEs. Statistical significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2: Descriptive Statistics: Teacher and External Observer Characteristics

	Mean	SD	N
Teacher Characteristics			
Female Teacher	0.70	0.46	41
Male Teacher	0.30	0.46	19
Teacher Age	42.37	10.71	53
EO Characteristics			
Female EO	0.75	0.43	7
Male EO	0.25	0.43	2

Note: Teacher information was collected during the baseline and endline surveys.

4 Results

4.1 Descriptive Statistics

In this section we document the summary statistics for the standardized *Pentabilities* scores from the classrooms recorded by four evaluator groups—self, peers, teachers, and EOs. For each group we compute the average rating at the behavior level and standardize it; we then aggregate behaviors into the five socioemotional domains[13], and finally re-standardize the resulting domain scores. Appendix A provides a detailed account of this procedure.

We begin with the self-assessments, disaggregated by gender and reported in Table B.14. There are no significant differences in how female and male students score themselves on *Autonomy*, *Cooperation*, and *Responsibility*. However, relative to their male classmates, female students rate themselves lower in *Emotional Management* and *Thinking Abilities*. The corresponding gender gaps—0.21 and 0.17 of a standard deviation (SD), respectively—are statistically significant according to unequal-variance t-tests reported in Table B.15.

Turning to peer evaluations, Table B.16 summarizes the scores classmates assign to one another. The first row of the table shows that on average, each student is evaluated by about seven

[13] As per the map in Figure 1.

different peers in the classroom, and there are 1,031 unique students in the peer-evaluation dataset. Panel A shows that, on average, female students receive higher peer scores than male classmates in every domain. Panels B and C, which disaggregate by peer evaluator gender, reveal a pronounced asymmetry. Female peers grade female classmates up and male classmates down, producing gender differentials between 0.29 and 0.46 SD, and all significant at the 1% level (Table B.17). Male peers also award higher scores to female students, but only in the domains of *Cooperation* and *Responsibility*, and the differences are smaller in magnitude. These patterns suggest that female peers are stricter toward male students and more favorable toward female students, particularly in *Autonomy*, *Cooperation*, and *Responsibility*, where the gaps exceed 40% of a SD .[14]

The descriptive evidence so far shows that female students are rated higher than males in peer-evaluations—a divergence mainly driven by female peers—even though females do not necessarily perceive themselves as better in these skills, as evident from their self-evaluations. We now compare these results with the teacher and EO-assigned scores in Tables B.18 and B.20. In the teacher dataset, each student is evaluated by an average of two teachers, with 1,136 students appearing in this sample. Female students receive higher scores overall relative to male students, chiefly driven by female teachers (all differences are significant at the 1% level; see Table B.19). Male teachers, in contrast, award significantly higher scores to female students only in the *Responsibility* domain.

As for the EOs, each observer evaluated 2-3 students in a given classroom, yielding a sample of 631 students.[15] Pooled across all observers, female students significantly lead male students in every domain except *Thinking Abilities*, as shown in Table B.21. Disaggregating by EO gender reveals further nuance. Female EOs give significantly higher scores to girls in three of five domains, though the magnitude is roughly half of that observed in teacher scores. Male EOs also award significantly higher scores to female students in four of five domains, including a 0.28 standard deviation gap in *Thinking Abilities*, where female EOs report no gender difference. The clearest asymmetries arise in *Autonomy* (male EOs give better scores to females) and *Emotional Manage-*

[14] To check whether the gender differences are present along the distribution of scores, we plot cumulative distribution functions (CDFs) for self and peer evaluations and report them in the Appendix B.1.

[15] Because only a subset of students in each classroom was randomly chosen for observation, the EO sample is roughly half the size of the teacher-evaluation sample.

ment (female EOs give higher score to females). Although the small number of male EOs (only two, versus seven female EOs) limits the precision of these estimates, the consistent premium for female students—especially from observers trained to be impartial and unfamiliar with the students—persists.

Overall, the descriptive evidence reveals systematic gender differences in socioemotional skill assessments. Female students tend to rate themselves more negatively than their male classmates in *Emotional Management* and *Thinking Abilities*. By contrast, peers, teachers, and external observers tend to evaluate girls more favorably in most domains. The divergence is largest in peer ratings, especially among female peers, who consistently assign higher scores to female classmates. Teacher and external observer assessments point in the same direction, with variation by evaluator gender and domain. Female teachers consistently assign higher scores to girls, whereas male teachers do so more selectively. External observers also rate girls higher, although the differences are smaller on average.

4.2 Empirical Strategy & Regression Results

The descriptive statistics already suggest a divergence along gender lines in socioemotional skill evaluations. In this section, we outline an empirical strategy to address our central research question: Are there significant gender differences in how youth evaluate themselves and their peers in terms of socioemotional skills?

4.2.1 Self-assigned Scores

We begin by defining our estimation strategy for self-assigned evaluations:

$$y_{is} = \alpha + \beta_1 \mathrm{StuFem}_i + \beta_2 Teacher_{is} + \beta_3 Peer_{is} + \beta_4 X'_{is} + \gamma + \varepsilon_{is}, \tag{1}$$

where $StuFem_i$ is a dummy variable that takes value 1 if the student is female. The term $Teacher_{is}$ represents the standardized teacher-assigned score for student i on skill domain s, averaged in cases where multiple teachers provided assessments. The inclusion of teacher-assigned scores controls for observed variation in student performance from the perspective of teachers. It does not measure students' underlying ability per se, but rather captures observable variation in skills as recognized by teachers. Similarly, we add the average peer-assigned score of student i, $Peer_{is}$, to control for the perception of the peers. X'_{is} denotes student background characteristics, including student birthplace, caregiver birthplace, and caregiver education levels. We include classroom fixed effects, γ ,to control for unobserved heterogeneity at the class level, and cluster standard errors at the student level, i.

Our empirical approach does not treat differences between self, peer, teacher, and external observer ratings as direct evidence of bias. Instead, we interpret divergences across these perspectives as reflecting systematic variation in how socioemotional skills are perceived and assessed. By triangulating across multiple evaluators, we are able to identify robust gendered patterns in the assessment and perception of adolescents' socioemotional skills.

We begin by estimating Equation 1 for the domain of *Autonomy*, with results presented in Table 3. The baseline specification controls for student background characteristics and classroom fixed effects only, while columns two and three progressively introduce controls for teacher-assigned and peer-assigned scores. Notably, the female student coefficient becomes more negative and statistically significant at the 10% level with these controls, suggesting that females underrate their *Autonomy* skill more critically once their skill level, as per their teachers and peers, has been partialled out. In other words, female students assign significantly lower scores to their *Autonomy* skill compared to male students, rating themselves approximately 14% of a SD lower after controlling for teacher and peer assessments.[16]

Table 3: Self-Assigned Score - Autonomy

	(1)	(2)	(3)
Female Student	-0.058	-0.141^{*}	-0.141^{*}
	(0.081)	(0.080)	(0.081)
Student's Teacher Score		0.301^{***}	0.177^{***}
		(0.053)	(0.061)
Ave Peer Score: autonomy			0.270^{***}
			(0.080)
Student Background Controls	*Yes*	*Yes*	*Yes*
Classroom FE	*Yes*	*Yes*	*Yes*
R-squared	0.127	0.185	0.207
Observations	570	540	525

Note: Robust standard errors in parentheses. * (p<0.10), ** (p<0.05), *** (p<0.01).

In Table 4, we present the full-model results for all five domains together.[17] The results in the last two columns reveal that, compared to their peers' and teachers' evaluations, female students underrate their *Emotional Management* and *Thinking Abilities* by 30% and 24% of a SD respectively, relative to male students. Notably, the average peer score shows a stronger and more significant correlation with self-scores than the teacher score. This closer alignment between peer and self evaluations may suggest that students share a common frame of reference when judging socioemotional competencies, one that differs from the teacher's perspective.[18]

[16] Expanded results with student background controls are reported in the Appendix Table C.22.

[17] For the remaining domains, we replicate the same regression analysis as in Table 3 and report the results in the Appendix (Tables C.23–C.26).

[18] As Table B.19 showed, female teachers tend to assign higher scores to female students. Therefore, as an additional check, we run the same regressions controlling teacher gender. The estimates are virtually unchanged, see Table C.27.

Table 4: Self-Assigned Score: All 5 dimensions

	Autonomy	Cooperation	Responsibility	Emotion	Thinking
Female Student	-0.141*	-0.101	-0.085	-0.302***	-0.239***
	(0.081)	(0.078)	(0.074)	(0.090)	(0.090)
Teacher Score	0.177***	0.117**	0.137***	0.087	0.042
	(0.061)	(0.058)	(0.053)	(0.058)	(0.066)
Average Peer Score	0.270***	0.290***	0.384***	0.319***	0.306***
	(0.080)	(0.074)	(0.068)	(0.089)	(0.080)
Observations	525	628	692	462	468
Adj. R-squared	0.207	0.211	0.231	0.191	0.249

Note: All regressions control for student background characteristics and classroom fixed effects. Robust standard errors are in parentheses. The dependent variable is self-assigned score. *p<0.10, ** p<0.05, *** p<0.010.

4.2.2 Peer-assigned Scores

Gender interaction regressions.

Next, we analyze peer-assigned evaluations using the following specification:

$$\begin{aligned} y_{ijs} &= \alpha + \beta_1 \text{StuFem}_{ij} + \beta_2 \text{PeerFem}_{ij} + \beta_3 \text{ StuFem} \times \text{PeerFem}_{ij} \\ &\quad + \beta_4 Self_{js} + \beta_5 Teacher_{is} + \beta_6 Self_{is} + \beta_7 Teacher_{js} \\ &\quad + \beta_8 X'_{is} + \beta_9 Z'_{js} + \gamma + \varepsilon_{ijs}, \end{aligned} \tag{2}$$

where *StuFem*$_{ij}$ is a dummy variable that takes value 1 if the student is female, (β_1), *PeerFem*$_{ij}$ indicates whether the peer evaluator is female (β_2), and their interaction $(StuFem \times PeerFem)_{ij}$ (β_3) captures any additional effect when both are female. $Self\,js$ and $Teacher_{js}$ control for the peer's own self- and teacher-assigned scores, while $Self_{is}$ and $Teacher_{is}$ account for the evaluated student's self- and teacher-assigned scores (in case of multiple teachers this is an average). $X'is$ and $Z'js$ include student and peer characteristics, with classroom fixed effects (γ) and clustering at the student, i, level.

Table 5 reports the peer-evaluation results for *Autonomy*. The coefficient on *StuFem*$_{ij}$ (β_1) reflects how female students are evaluated relative to male students by male peers; *PeerFem*$_{ij}$ (β_2) indicates the impact of having a female peer evaluator; and the interaction term (β_3) captures any additional effect when both the student and evaluator are female. The net effect of a female peer evaluating a female versus a male student ($\beta_1 + \beta_3$) is reported at the bottom of each table.

The specification evolves across columns. Column (1) includes background controls and classroom fixed effects. Column (2) adds the peer's self score to assess whether self-perceptions shape peer evaluations. Column (3) introduces the teacher-assigned score for the student being evaluated, isolating variation in skill as recognized by teachers. Column (4) then adds the student's self-assigned score, allowing us to control for the evaluated student's own self-perception. Finally, Column (5) incorporates the peer's teacher-assigned domain score, which enables us to test whether peers who are more highly rated by teachers themselves evaluate others differently.

In the first column of Table 5, the coefficient on β_1 suggests that male peers do not evaluate female students any differently, however the negative and significant female peer coefficient (β_2) suggests that female peers give 0.21 of a SD lower scores to male students. When the student is female, this negative effect flips, as shown by the interaction term. The net effect, given by the term $\beta_1 + \beta_3$ below, suggests that female peers give almost half a SD higher *Autonomy* scores to female students compared to the male students.

The second column shows a positive and significant relationship with Peer's Self Score, suggesting that peers who see themselves as more autonomous also tend to evaluate others more favorably. While including this variable does not alter the main gender interaction, adding the Student's Teacher Score in Column (3) attenuates the female–female premium to roughly 0.33 SD. Similarly, the net effect remains positive and significant when we add Student's Self Score in the next column, indicating that while part of the advantage is explained by differences in the skill level (to the extent that it is captured by the teacher and self assessments), a notable positive female–female premium persists. Finally, in Column (5), the negative coefficient on the Peer's Teacher Score suggests that peers with stronger teacher-assessed abilities are somewhat harsher graders (although smaller in magnitude). Overall, these results highlight a robust female–female advantage in *Autonomy* evaluations, and the relevance of both self-perception and teacher assessments in shaping peer-graded outcomes.

Table 5: Peer-Assigned Score - Autonomy

	(1)	(2)	(3)	(4)	(5)
Female Student (β_1)	0.034	0.050	−0.082	−0.088	−0.086
	(0.069)	(0.072)	(0.065)	(0.065)	(0.065)
Female Peer (β_2)	-0.212^{***}	-0.210^{***}	-0.225^{***}	-0.235^{***}	-0.207^{***}
	(0.065)	(0.067)	(0.063)	(0.064)	(0.064)
Female Student x Female Peer (β_3)	0.438^{***}	0.423^{***}	0.416^{***}	0.446^{***}	0.454^{***}
	(0.084)	(0.086)	(0.082)	(0.083)	(0.084)
Peer's Self Score		0.145^{***}	0.154^{***}	0.167^{***}	0.183^{***}
		(0.026)	(0.025)	(0.025)	(0.026)
Student's Teacher Score			0.416^{***}	0.361^{***}	0.364^{***}
			(0.034)	(0.038)	(0.038)
Student's Self Score				0.121^{***}	0.116^{***}
				(0.040)	(0.040)
Peer's Teacher Score					-0.083^{***}
					(0.028)
Fem Peer: Fem vs Male Stu ($\beta_1 + \beta_3$)	0.472	0.472	0.334	0.358	0.368
Fem Peer: P-val	0.000	0.000	0.000	0.000	0.000
Student Background Controls	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Peer Background Controls	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Classroom FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
R-squared	0.170	0.191	0.283	0.281	0.286
Observations	2420	2232	2184	2038	2006

Standard errors in parentheses and clustered at student level. * (p<0.10), ** (p<0.05), *** (p<0.01).

We replicate the same procedure to construct the tables for the remaining domains, with detailed results from all specifications, including control variables, presented in Tables C.28–C.32 of the Appendix. Table 6 reports the full models for all domains. Across all five outcomes, the coefficient on β_2 is consistently negative and statistically significant, indicating that female graders assign notably lower scores to male students. In contrast, the β_1 coefficient is generally small or insignificant, implying that male peers do not systematically differentiate between male and female students. The interaction term β_3 is large and highly significant for each outcome, yielding a net positive female–female effect ($\beta_1 + \beta_3$) of 0.21 to 0.37 SD (all $p < 0.01$) for these constructs—underscoring that, in each dimension, female students receive substantially higher ratings from female peers.

In addition to these gender effects, we observe that a higher Peer Self Score is systematically associated with more generous evaluations, with effect sizes ranging from 0.18 to 0.26 SD. Student's Teacher Score is likewise positive in each column—magnitudes span 0.28 to 0.42 SD—suggesting that stronger teacher-assessed performance translates into higher peer-assigned ratings. Student's Self Score also matters, but to a lesser degree (0.09 to 0.17 SD). In contrast, the coefficient on

Peer's Teacher Score is negative and significant in four of the five domains, implying that peers who are highly rated by teachers often apply a stricter grading standard to others.

Taken together, these findings reinforce two main insights: (i) a consistent, domain-spanning female–female advantage, and (ii) the significant role of both self-perceptions and teacher evaluations in shaping peer assessments. Given that teachers, especially female teachers, tend to assign higher scores to female students, controlling for teacher-assigned scores in our peer regressions isolates the peer premium that is independent of teacher recognition.[19] As a result, our estimates likely provide a conservative (lower-bound) measure of the female–female peer premium, reflecting only the incremental effect above and beyond what is captured by teacher evaluations.

Table 6: Peer-Assigned Scores: All 5 dimensions

	Autonomy	Cooperation	Responsibility	Emotion	Thinking
Female Student (β_1)	-0.086	-0.112**	-0.019	-0.044	-0.110
	(0.065)	(0.056)	(0.056)	(0.059)	(0.069)
Female Peer (β_2)	-0.207***	-0.293***	-0.156***	-0.231***	-0.205***
	(0.064)	(0.053)	(0.055)	(0.056)	(0.067)
Female Student x Female Peer (β_3)	0.454***	0.406***	0.232***	0.357***	0.415***
	(0.084)	(0.068)	(0.070)	(0.076)	(0.081)
Peer's Self Score	0.183***	0.264***	0.205***	0.192***	0.250***
	(0.026)	(0.021)	(0.021)	(0.020)	(0.026)
Student's Teacher Score	0.364***	0.349***	0.418***	0.281***	0.370***
	(0.038)	(0.033)	(0.030)	(0.030)	(0.036)
Student's Self Score	0.116***	0.091***	0.094***	0.102***	0.172***
	(0.040)	(0.029)	(0.030)	(0.030)	(0.031)
Peer's Teacher Score	-0.083***	0.019	-0.061**	0.081***	-0.145***
	(0.028)	(0.027)	(0.025)	(0.025)	(0.030)
$\beta_1 + \beta_3$	0.368***	0.294***	0.213***	0.312***	0.305***
Observations	2006	2842	2762	2239	2065
Adj. R-squared	0.263	0.249	0.242	0.217	0.284

Note: All regressions control for student and peer background characteristics, and classroom fixed effects. Standard errors clustered at student level in parentheses. The dependent variable is peer-assigned score. *p<0.10, ** p<0.05, *** p<0.01.

Fixed-effect regressions.

The results, thus far, suggest a systematic female–female premium across all socioemotional domains, even though females do not rate themselves higher in any of these areas, and in fact evaluate themselves significantly lower in three of the domains. However, several unobserved factors may

[19] We also run the same regressions controlling teacher gender and find that the main estimates stay the same, see Table C.33.

still confound these estimates. These include differences in how peers engage with educational technology (edtech), idiosyncrasies in grading styles, and the unique ways students demonstrate socioemotional skills in the classroom. To address these concerns, we introduce peer and student fixed effects (FEs). This strategy enables two complementary within-person comparisons. First, Peer FEs compare how the same peer evaluator (female or male) assigns scores to male versus female students. Second, Student FEs compare how the same student is graded by female versus male peers, controlling for any time-invariant student characteristics.

Following Boring (2017), we estimate the following Peer and Student FE models:

$$\begin{aligned} y_{ijs} &= \alpha + \beta_1 \, \text{StuPeerFem}_{ij} + \beta_2 \, \text{StuPeerMale}_{ij} + \beta_3 Teacher_{is} + \beta_4 Self_{is} + \beta_5 X'_{is} \\ &\quad + \nu_j + \varepsilon_{ijs} \end{aligned} \tag{3}$$

$$\begin{aligned} y_{ijs} &= \alpha + \beta_1 \, \text{StuPeerFem}_{ij} + \beta_2 \, \text{StuPeerMale}_{ij} + \beta_3 Teacher_{js} + \beta_4 Self_{js} + \beta_5 Z'_{js} \\ &\quad + \eta_i + \varepsilon_{ijs} \end{aligned} \tag{4}$$

where y_{ijs} is the peer-assigned score student i receives from peer j on domain s. The key variables of interest are $StuPeerFem_{ij}$ and $StuPeerMale_{ij}$, which take the value 1 if a female peer is grading a female student or a male peer is grading a male student, respectively.

Under the Peer FEs model (Equation 3), ν_j absorbs all time-invariant characteristics of peer j. The variables $Teacher_{is}$ and $Self_{is}$ represent the student's teacher-assigned and self-assigned scores, and X'_{is} is the student's background characteristics. In the Student FEs model (Equation 4), η_i absorbs all time-invariant characteristics of student i, and the variables $Teacher_{js}$ and $Self_{js}$ reflect the peer's own teacher- and self-assigned scores. Z'_{js} represents peer-specific background information.

Table 7 presents results from the Peer and Student FE regressions for all 5 domains. In the Peer FEs columns, the coefficient on *Fem Student x Female Peer* captures how the same female peer scores female versus male students, while *Male Student x Male Peer* shows the differential scoring by the same male peer.[20] Since we are controlling for peer's fixed effects, teacher score, self score and background characteristics all refer to the student. In the Student FEs column, we examine how the same student is rated by peers of different genders. Here, teacher and self scores—as well as background controls—correspond to the evaluating peer.

Across all domains, the *Fem Student x Fem Peer* coefficient under Peer FEs shows a robust female–female premium in the range of 26%-32% of a SD, which is both statistically and economically significant. In contrast, male peers do not systematically favor male students: the *Male Student x Male Peer* coefficient remains small and insignificant in all domains under the Peer FEs specification. Turning to the Student FEs columns, we find generally consistent evidence. With the exception of *Responsibility*, female students receive higher scores from female than male peers. For male students, male peers tend to award higher scores (relative to the female peers, who favor female students), which is reflected in positive and significant coefficients in all domains except *Autonomy*.

Regarding the control variables, Teacher Score variable under Peer FE suggests that peers typically award higher scores to students who also receive strong teacher ratings. Students' self-scores correlate positively only in *Responsibility*, *Emotional Management* and *Thinking Abilities*, with a smaller effect for *Cooperation*. In the Student FEs model—where teacher and self scores refer to peer characteristics—we observe that peers with higher teacher-assigned scores grade more stringently in *Autonomy* and *Thinking Abilities*, suggesting they apply a stricter or harsher grading standard. Meanwhile, peers with stronger self-evaluations tend to give higher marks overall.

To summarize, the fixed-effects models provide several nuanced insights. Most notably, a substantial female–female premium in peer assessments persists across all socioemotional domains, even after accounting for unobserved heterogeneity in grading styles and student characteristics. This premium, which ranges from 26% to 32% SD, remains statistically significant and consistent across domains. In contrast, male–male peer effects are generally small and not consistently

[20] This specification is equivalent to running two separate regressions, each including only one of the interaction terms (e.g., *Fem Student x Female Peer*), while replacing the other interaction with the corresponding gender dummy variable. For illustration, the results from these separate estimations for the domain of *Autonomy* are reported in the Appendix Table C.34.

distinguishable from zero, suggesting that male peers do not systematically differentiate in their evaluations by student gender. Turning to the control variables, the fixed-effects regressions reinforce the earlier finding that students rated more highly by teachers also tend to receive higher peer scores, while peers with stronger self-evaluations typically provide more generous ratings. There is also some indication that peers who themselves are highly rated by teachers may apply stricter grading standards, particularly in *Autonomy* and *Thinking Abilities*.

Table 7: Peer-Assigned Score: Peer FE vs. Student FE Across 5 Domains

	Autonomy		**Cooperation**		**Responsibility**		**Emotion Mngt.**		**Thinking Ab.**	
	Peer FE	**Student FE**	**Peer FE**	**Student FE**	**Peer FE**	**Student FE**	**Peer FE**	**Student FE**	**Peer FE**	**Student FE**
Fem Student x Fem Peer	0.311***	0.246***	0.267***	0.161***	0.257***	0.073	0.301***	0.131**	0.318***	0.218***
	(0.074)	(0.063)	(0.058)	(0.051)	(0.061)	(0.052)	(0.069)	(0.056)	(0.080)	(0.052)
Male Student x Male Peer	0.004	0.127	0.045	0.200***	-0.030	0.118*	-0.036	0.163***	-0.002	0.218***
	(0.076)	(0.078)	(0.061)	(0.060)	(0.063)	(0.061)	(0.062)	(0.056)	(0.076)	(0.073)
Self Score	0.076	0.170***	0.061*	0.250***	0.078**	0.206***	0.097***	0.195***	0.172***	0.220***
	(0.047)	(0.031)	(0.032)	(0.023)	(0.034)	(0.024)	(0.035)	(0.022)	(0.039)	(0.028)
Teacher Score	0.378***	-0.086***	0.378***	0.009	0.422***	-0.041	0.266***	0.065**	0.350***	-0.172***
	(0.047)	(0.033)	(0.037)	(0.029)	(0.033)	(0.027)	(0.033)	(0.025)	(0.041)	(0.032)
Peer Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Student Controls	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
Observations	2297	2405	3264	3363	3216	3382	2725	2897	2498	2612
Adj. R-squared	0.518	0.573	0.492	0.504	0.492	0.539	0.485	0.465	0.518	0.526

Note: Standard errors in parentheses and clustered at the student level. The dependent variable is peer-assigned score. Under Peer FEs, the Teacher and Self Score refers to student's scores and under Student FEs, these refer to the peer's scores . *p<0.10, ** p<0.05, *** p<0.010.

5 Robustness Checks

We conduct three sets of robustness checks to assess the validity of our main findings. First, we re-estimate the primary specifications for self and peer evaluations, substituting teacher-assigned socioemotional scores with those recorded by EOs within the classroom setting. Second, we replicate the analysis using data from the standardized "lab-in-the-field" activity, which was conducted outside the regular classroom context and monitored exclusively by EOs. Third, we use a more conventional survey instrument to measure socioemotional skills, collected at endline as part of the study's questionnaire battery, to examine whether the patterns documented with the behavior-based instrument in the classroom and the lab activity are reflected in an alternative, survey-based measurement framework.

5.1 Classroom Observations EO scores

We begin with the self-assigned scores. Table 8 re-runs the self-evaluation regressions using the EO scores instead of the teacher scores. Because EOs assessed only a subsample of students, the estimation sample is approximately half the size of the baseline. Despite this loss of precision, the main pattern remains: the female coefficient is still negative, though no longer statistically significant (except slightly in *Responsibility*). This attenuation arises because EOs assign a smaller premium to female students compared to teachers, narrowing the gap between girls' self-assessments and external evaluations. As a result, the gender gap in self-reported evaluations diminishes, mirroring the smaller raw gender differences documented in Table B.21.

Turning to peer evaluations, Table 9 shows that the sample shrinks to about one-quarter of the original size, yet the main findings persist. Female peers continue to rate female classmates more favorably in every domain and the net effect, given by $\beta_1 + \beta_3$, ranges between 0.20-0.40 SD. Moreover, a student's own EO score is positively associated with the peer-assigned score, while the peer's EO score is mostly not, replicating the baseline pattern.[21]

[21] FE regressions in Appendix Table C.35 show that the female interaction remains positive and significant in all domains under the peer fixed effects, indicating a persistent female advantage in the range of 0.28-0.46 SD.

Taken together, substituting teacher evaluations with a more objective external benchmark, the EO scores, does not alter our findings. If anything, these results reinforce the conclusion that the observed female premium in socioemotional skills is robust to the use of alternative classroom observers.

Table 8: Robustness Check with EO Scores - Self-Assigned Score: All 5 dimensions

	Autonomy	Cooperation	Responsibility	Emotion	Thinking
Female Student	-0.086	-0.123	-0.177*	-0.182	-0.184
	(0.116)	(0.118)	(0.104)	(0.121)	(0.143)
EO Score	0.043	0.108	0.092	0.035	0.078
	(0.072)	(0.088)	(0.082)	(0.080)	(0.092)
Average Peer Score	0.355***	0.227*	0.528***	0.515***	0.453***
	(0.109)	(0.119)	(0.091)	(0.130)	(0.113)
Observations	260	301	365	234	206
R-squared	0.345	0.327	0.394	0.331	0.423

Note: All regressions control for student background characteristics and classroom fixed effects. Robust standard errors are in parentheses. The dependent variable is self-assigned score. *p<0.10, ** p<0.05, *** p<0.010.

Table 9: Robustness Check with EO Scores - Peer-Assigned Scores: All 5 dimensions

	Autonomy	Cooperation	Responsibility	Emotion	Thinking
Female Student (β_1)	-0.230*	-0.053	0.112	0.078	-0.209
	(0.133)	(0.103)	(0.096)	(0.126)	(0.177)
Female Peer (β_2)	-0.291**	-0.461***	-0.192*	0.023	-0.261*
	(0.122)	(0.109)	(0.102)	(0.132)	(0.155)
Female Student x Female Peer (β_3)	0.511***	0.453***	0.090	0.216	0.518**
	(0.173)	(0.137)	(0.129)	(0.185)	(0.244)
Peer's Self Score	0.202***	0.239***	0.229***	0.305***	0.258***
	(0.045)	(0.044)	(0.046)	(0.053)	(0.066)
Student's EO Score	0.216***	0.119**	0.235***	0.231***	0.164**
	(0.070)	(0.060)	(0.060)	(0.050)	(0.079)
Student's Self Score	0.168**	0.138***	0.114**	0.105**	0.202***
	(0.071)	(0.048)	(0.049)	(0.050)	(0.064)
Peer's EO Score	-0.071	0.132**	0.009	-0.027	0.004
	(0.046)	(0.056)	(0.047)	(0.059)	(0.067)
$\beta_1 + \beta_3$	0.280**	0.401***	0.202*	0.294**	0.309*
Observations	445	573	696	410	311
Adj. R-squared	0.234	0.233	0.215	0.231	0.268

Note: All regressions control for student and peer background characteristics, and classroom fixed effects. Standard errors clustered at student level in parentheses. The dependent variable is peer-assigned score. *p<0.10, ** p<0.05, *** p<0.01.

5.2 Lab-in-the-field Activity

We next assess the robustness of our findings using data from a lab-in-the-field exercise conducted outside regular classroom hours. This activity was carried out in a standardized, controlled environment, with random assignment of students to groups and no academic stakes, thereby mitigating potential confounds such as teacher-student relationships, pre-existing peer dynamics, or endogenous group gender composition. Immediately after the exercise, both students and EOs independently completed a 10-item questionnaire that maps directly onto the five-domain socioemotional skills framework used in our main analysis. Each rater evaluated how well participants demonstrated specific behaviors during the activity, with two questions corresponding to each of the five socioemotional domains, using a 5-point scale.[22] Descriptive statistics for each domain and for each type of rater are provided in Appendix Tables C.37–C.42.

As shown in Table 10, even in this different setting, female students continue to rate themselves more critically, conditional on peer and EO scores. The female coefficient remains negative and

[22] Unlike the Pentabilities measure, these ratings are not standardized for the analysis.

statistically significant for *Emotion Management* ($p < 0.10$) and *Thinking Abilities* ($p < 0.05$), while differences in other domains are not statistically distinguishable from zero. Importantly, peer score and EO-assigned scores are strong positive predictors of students' self-evaluations in every domain, indicating that self-assessments remain closely anchored to externally observable performance. Yet, females persistently downplay their performance precisely in those domains where classroom-based analyses previously revealed significant gender gaps.

Table 10: Lab-in-the-field: Self-Assigned Score

	Autonomy	Cooperation	Responsibility	Emotion	Thinking
Female Student	-0.047	0.020	0.006	-0.118*	-0.164**
	(0.054)	(0.059)	(0.056)	(0.064)	(0.065)
EO Score	0.253***	0.128***	0.144***	0.164***	0.168***
	(0.035)	(0.045)	(0.037)	(0.047)	(0.049)
Average Peer Score	0.232***	0.253***	0.219***	0.217***	0.270***
	(0.051)	(0.050)	(0.051)	(0.054)	(0.056)
Male Student Mean	3.84	3.88	4.12	3.77	3.74
Observations	769	768	770	766	767
Adj. R-squared	0.204	0.121	0.145	0.103	0.108

Note: All regressions control for student background characteristics and classroom fixed effects. Robust standard errors are in parentheses. The dependent variable is self-assigned score. *p<0.10, ** p<0.05, *** p<0.010.

Turning to peer evaluations, Table 11 echoes our central classroom finding: female peers systematically rate their female classmates more favorably. Specifically, $\beta_1 + \beta_3$—the net difference in how female peers evaluate female versus male classmates—we find a female premium ranging from 0.12 to 0.27, statistically significant at the 5-percent level, or less, in every domain. These regressions also reveal that a peer's self-assigned score is a strong predictor of how they rate others, suggesting that individuals' self-perceptions shape their evaluations of peers. Consistent with earlier results, a student's EO and self scores add additional predictive power, though to a lesser extent. Notably, peers with higher EO scores tend to provide somewhat harsher ratings in several domains, reinforcing previous findings that stronger performers may hold their peers to higher standards.[23]

[23] We also run fixed-effects robustness checks on peer-assigned scores from the lab-in-the-field in Appendix Table C.36 and find that female evaluators consistently give higher scores to female classmates, yielding a substantial advantage (0.18 to 0.33) in all domains except *Emotion Management*.

Table 11: Lab-in-the-field: Peer-Assigned Scores

	Autonomy	Cooperation	Responsibility	Emotion	Thinking
Female Student (β_1)	-0.005	0.067	0.084	0.052	-0.022
	(0.058)	(0.056)	(0.057)	(0.060)	(0.060)
Female Peer (β_2)	0.043	-0.116**	-0.110*	0.003	-0.074
	(0.059)	(0.056)	(0.058)	(0.060)	(0.060)
Student x Peer Female (β_3)	0.129*	0.154**	0.187**	0.077	0.220***
	(0.078)	(0.077)	(0.077)	(0.082)	(0.080)
Peer's Self Score	0.405***	0.503***	0.462***	0.355***	0.392***
	(0.030)	(0.030)	(0.032)	(0.029)	(0.026)
Student's EO Score	0.287***	0.236***	0.271***	0.199***	0.262***
	(0.027)	(0.032)	(0.034)	(0.033)	(0.034)
Student's Self Score	0.167***	0.113***	0.120***	0.101***	0.143***
	(0.034)	(0.032)	(0.035)	(0.028)	(0.029)
Peer's EO Score	-0.076***	-0.005	-0.074**	0.027	-0.083***
	(0.025)	(0.027)	(0.029)	(0.030)	(0.030)
$\beta_1 + \beta_3$	0.124**	0.221***	0.271***	0.129**	0.198***
Male × Male Mean	3.81	3.79	3.88	3.73	3.79
Observations	1859	1853	1858	1828	1841
R-squared	0.352	0.380	0.363	0.293	0.325

Note: All regressions control for student and peer background characteristics, and classroom fixed effects. Standard errors clustered at student level in parentheses. The dependent variable is peer-assigned score. *p<0.10, ** p<0.05, *** p<0.01.

Taken together, the lab-in-the-field evidence strongly reinforces our main findings. Even after removing classroom-specific dynamics, potential teacher biases, and friendship-driven influences, the key asymmetry persists: female peers systematically recognize and reward the socioemotional strengths of their female classmates, despite their critical self-assessments. These patterns, first documented through extensive classroom evaluations over six months, reappear with clarity in a simplified context using only a brief, standardized 10-item survey.

5.3 BESSI

The previous section showed that a concise survey can effectively capture the underlying gender dynamics. A natural next step is to ask whether a more conventional, validated survey of socioemotional skills yields the same conclusions. To do so, we turn to a measurement tool called the BESSI, which organizes skills into five domains—*Self-Management*, *Social Engagement*, *Cooperation*, *Emotional Resilience*, and *Innovation*. Skills are conceptualized as functional capacities: how well individuals can enact a behavior when needed, rather than how often they do so.[24] The instrument has been validated for adolescents and adults, in both self- and observer-report, with high internal consistency at domain and facet level, strong convergent–discriminant validity, and incremental predictive power beyond the Big Five. As with Pentabilities, BESSI yields a 360-degree view by combining self, peer, and teacher reports. At endline, students completed the 20-item self-report and answered 20-item BESSI observer-reports for two classmates; teachers also completed 20-item BESSI observer-reports for a subsample of students.[25] Items are scored on a 1–5 scale; domain scores average the relevant items.[26]

Table 12 reports self-reported BESSI scores. Conditioning on student covariates, classroom fixed effects, and the student's BESSI teacher and average peer scores, female students rate themselves significantly lower in four of five domains (effects between 0.16 and 0.47 points on the 1–5

[24] For example: "How well can you plan out your time?"

[25] Short-form instruments and scoring instructions, including the BESSI-20, are available at the SEB Skills Lab website.

[26] Descriptive statistics for each domain and type of rater are provided in Appendix Table C.43.

scale), with no significant difference for *Innovation*.[27] This pattern aligns with the main results: girls are more self-critical, particularly in emotional management, while also underscoring that the the gap is not universal across domains.

Table 13 turns to peer-assigned BESSI scores. Female peers assign lower scores on average to male students, but the *Female Student* × *Female Peer* interaction is positive and highly significant in all domains. The implied female–female premium, $\beta_1 + \beta_3$, is consistently significant and in the range of 0.26–0.36 points on the 1–5 scale. As in our baseline specifications, a student's BESSI teacher score and the peer rater's own BESSI self-score strongly predict assigned ratings; the student's self-score is also positively associated in most domains.[28]

In summary, the three robustness exercises—substituting teacher evaluations with EO ones in the classroom, replicating the analysis in a standardized lab-in-the-field setting, and introducing a validated survey inventory—support the main results. Female students' self-assessments are systematically lower, while peers and other observers evaluate them more favorably; peer ratings feature a pronounced female-female premium. Taken together, this triangulation across observers (self, peer, teacher, EO), settings (classroom, lab-in-the-field), and methods (behavioral observations, survey inventory) strengthens the finding that the gender asymmetries we document are robust across different contexts and instruments.

[27] *Innovation* in BESSI (openness to new ideas, creativity, exploratory problem-solving) maps closely to Pentabilities' *Thinking Abilities* (reflection and strategic problem-solving). Both capture the cognitive dimension of SEL; terminology differs, but the constructs are closely aligned.

[28] Given time constraints, teachers could only complete BESSI observer-reports for a subsample of students at the endline data collection. Therefore, there is less of an overlap and adding the peer rater's teacher-assigned BESSI score reduces the estimation sample substantially. However, as Table C.49 in Appendix shows, the results are unchanged in that specification.

Table 12: BESSI Self-Assigned Score (Endline): All 5 dimensions

	Self Mngt	Social Engt	Cooperation	Emotional Mngt	Innovation
Female Student	-0.229***	-0.309***	-0.162***	-0.467***	0.060
	(0.064)	(0.069)	(0.060)	(0.073)	(0.069)
Teacher Score	0.199***	0.298***	0.169***	0.180***	0.155***
	(0.047)	(0.051)	(0.053)	(0.062)	(0.060)
Average Peer Score	0.071	0.148**	0.132**	0.071	0.139**
	(0.073)	(0.072)	(0.063)	(0.081)	(0.067)
Male Student Mean	3.52	3.31	3.72	3.42	3.30
Observations	510	509	509	509	510
Adj. R-squared	0.079	0.146	0.065	0.097	0.056

Note: All regressions control for student background characteristics and classroom fixed effects. Robust standard errors are in parentheses. The dependent variable is self-assigned score. *p<0.10, ** p<0.05, *** p<0.010.

Table 13: BESSI Peer-Assigned Score (Endline): Specification without Peer's Teacher Score

	Self Mngt	Social Engt	Cooperation	Emotional Mngt	Innovation
Female Student	-0.005	-0.021	0.026	0.030	0.120**
	(0.059)	(0.057)	(0.056)	(0.062)	(0.056)
Female Peer	-0.366***	-0.202***	-0.194***	-0.241***	-0.162***
	(0.056)	(0.056)	(0.057)	(0.058)	(0.056)
Female Student × Female Peer	0.363***	0.278***	0.263***	0.334***	0.240***
	(0.076)	(0.084)	(0.079)	(0.077)	(0.076)
Peer's Self Score	0.230***	0.231***	0.392***	0.241***	0.241***
	(0.033)	(0.030)	(0.035)	(0.031)	(0.029)
Student's Teacher Score	0.438***	0.303***	0.337***	0.345***	0.443***
	(0.026)	(0.031)	(0.033)	(0.033)	(0.030)
Student's Self Score	0.062	0.101***	0.132***	0.101***	0.122***
	(0.038)	(0.034)	(0.039)	(0.032)	(0.030)
$\beta_1 + \beta_3$	0.358***	0.257***	0.289***	0.364***	0.360***
Male × Male Mean	3.32	3.37	3.44	3.36	3.18
Observations	1292	1290	1289	1288	1288
Adj. R-squared	0.391	0.233	0.274	0.248	0.310

Note: All regressions control for student and peer background characteristics, and classroom fixed effects. Standard errors are in parentheses and clustered at student level. The dependent variable is peer-assigned score. *p<0.10, ** p<0.05, *** p<0.01.

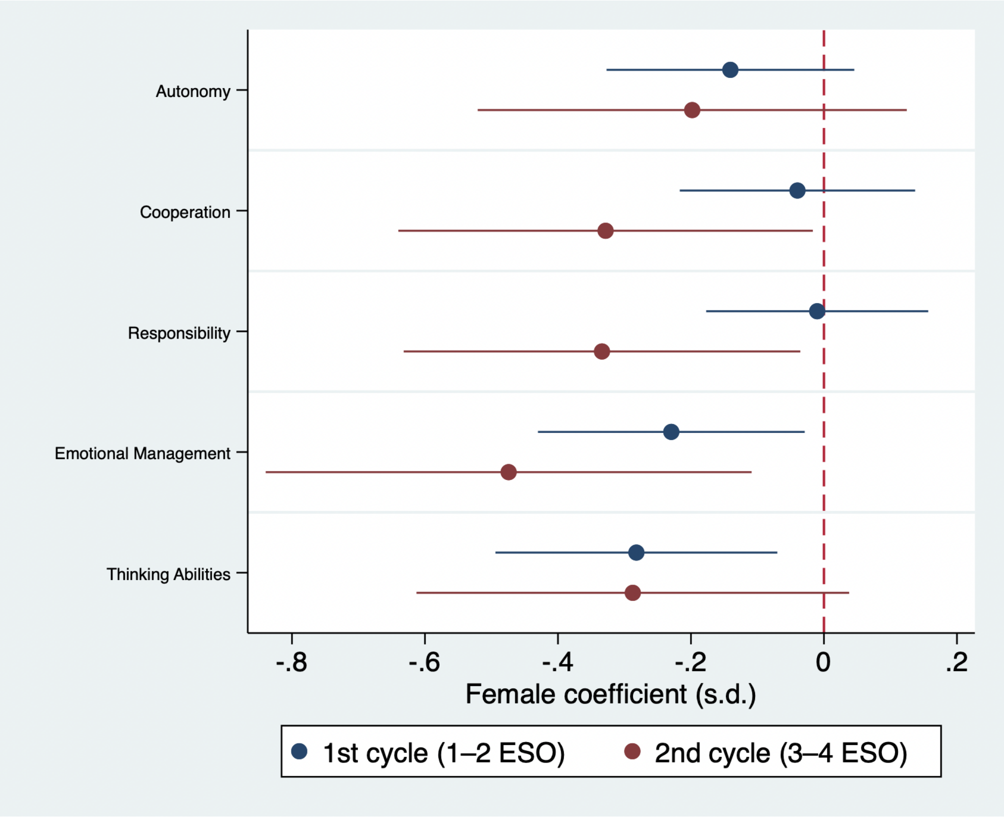
6 Heterogeneity

6.1 ESO Levels

The evidence thus far documents a robust gender gap in both self- and peer-assessments of socioemotional skills, but the magnitude and nature of these gaps may evolve as students advance through adolescence. Early secondary school students—those in the first cycle of ESO (ages 12–14)—are likely to differ from older cohorts both in their self-perceptions and in how they assess their peers, reflecting the rapid development of self-concept and the shifting dynamics of adolescent social networks. To capture this heterogeneity, we re-estimate our main specifications separately for Cycle 1 (1st–2nd ESO) and Cycle 2 (3rd–4th ESO).

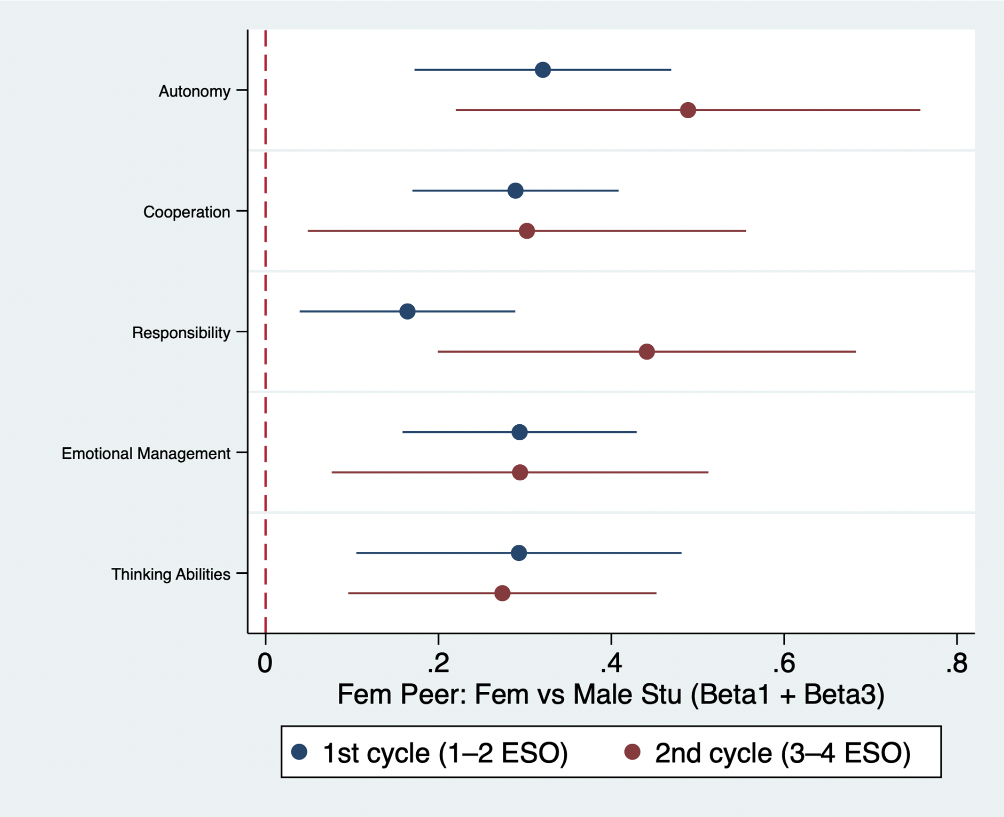
Figure 2 plots the female-student coefficient for each socioemotional domain by cycle. A value of zero indicates parity with boys; negative values denote lower self-ratings for girls. The age gradient is pronounced in the figure. In Cycle 1, girls rate themselves lower in *Emotional Management* (-0.23 SD) and *Thinking Abilities* (-0.25 SD), with negligible gaps in *Cooperation* and *Responsibility*. In Cycle 2, the landscape shifts: the gap in *Emotional Management* nearly doubles (-0.44 SD), and sizable and significant deficits emerge in *Cooperation* (-0.33 SD) and *Responsibility* (-0.31 SD), while the difference in *Thinking Abilities* disappears. These results suggest that, as students progress through secondary school, gender disparities in self-assessment become more pronounced in domains involving social interaction and affect regulation, but narrow in domains more closely tied to cognitive ability.

Figure 2: Female Self-Assigned Score by Cycle



Turning to peer assessments, Figure 3 reports $\beta_1 + \beta_3$—the additional score a female receives when evaluated by a female peer rather than a male classmate—from Model 2, estimated separately for each cycle. The heterogeneity analysis reveals that the female-in-group premium varies by domain and age group. In *Autonomy*, the premium is positive and significant in both cycles, and larger among older pupils (0.30 SD in Cycle 1, 0.44 SD in Cycle 2). For *Cooperation*, *Emotional Management*, and *Thinking Abilities*, the premium is present in Cycle 1 but fades in Cycle 2. For *Responsibility*, the premium is negligible in the first cycle but becomes sizable and significant in the second, uniquely exhibiting a stronger positive effect for the older cohort. Thus, while a female-female premium is evident, its magnitude depends both on the skill under scrutiny and the stage of schooling.[29]

Figure 3: Female Peer Effect by Cycle



[29] We also examined heterogeneity in the deviation between self and peer-assigned scores across several factors (i.e., teacher/EO score, emotional-intelligence as measured by the Eyes Test, socioeconomic background, and teacher characteristics) and found no systematic patterns.

7 Discussion and Conclusion

Persistent gender disparities in labor market outcomes underscore the importance of subtler mechanisms, particularly the role of perceptions and beliefs about individual abilities. Identifying when gender differences in beliefs and the evaluations they shape first emerge, and how best to mitigate them, is essential for policy design. This paper advances our understanding by documenting gender gaps in evaluations of socioemotional skills in early adolescence, leveraging a rich dataset that combines evaluations by students themselves, their peers, teachers, and external observers. Triangulating these perspectives, the study reveals systematic gendered patterns in both self-assessments and peer evaluations, providing a basis for targeted interventions and future research.

We document three central findings. First, female students exhibit a robust and negative gap in self-evaluations, especially in domains culturally coded as masculine (Emotion Management and Thinking Abilities). This finding resonates with prior literature highlighting women's systematic undervaluation of their abilities in tasks associated with male stereotypes. Given the well-documented links between self-perceptions, educational investment, and occupational choice, such critical self-assessments are likely to generate downstream disparities in track choice, field of study, and the types of careers girls consider attainable.

Second, peer evaluations display a distinctive asymmetry. Conditional on teacher and external-observer scores, female students assign systematically higher socioemotional scores to their female classmates in every domain. We interpret this female–female premium as evidence that adolescent girls recognize and affirm strengths in their female peers that they do not attribute to themselves. Importantly, given the lower self-assessments documented earlier, this positive peer recognition seems to be driven by perceptions held by others rather than self-promotion. This fresh empirical evidence expands our understanding of gendered evaluation dynamics among adolescents, highlighting a dual dynamic of negative self-assessment alongside affirmative peer recognition.

Third, these gendered evaluation patterns are highly robust. They remain when teacher ratings are replaced by external-observer scores, and they replicate in both a lab-in-the-field exercise and the endline socioemotional skills inventory. Their persistence across rater types, settings, and measurement methods suggests that we are capturing a broad and stable feature of how socioemotional competence is perceived in early adolescence, with relevance that extends beyond the classroom.

Collectively, the findings indicate that gendered patterns of ability perception crystallize in early adolescence, precisely when self-beliefs and aspirations remain malleable (Steinberg, 2014). This developmental window offers an opportunity to mitigate emerging gaps, and schools are a natural platform for scalable interventions. Existing evidence shows that socioemotional skills respond to structured school-based programs and that such programs can correct misperceptions and reduce bias (Alan, 2025). Because the school environment is a central part of adolescents' social networks, shifting perceived norms through visible peer referents can move behavior at relatively low cost. In related work, classroom discussions that explicitly interrogate gender beliefs have produced durable changes in attitudes among similar age groups, with effects persisting two years after program completion (Dhar et al., 2022), while brief coeducational discussion groups that surface misperceived norms around masculinity have reduced misperceptions and increased willingness to engage (Matavelli, 2025).

A practical implication would be to provide individualized feedback that makes peer-recognized strengths salient to girls, in order to counteract overly critical self-assessments. Making explicit which socioemotional strengths classmates consistently recognize may help close the gap between self- and peer beliefs, support more accurate self-views, and ultimately influence educational and occupational choices. These ideas resonate with evidence on the importance of relatable female role models (Porter and Serra, 2020) and with work showing that socioemotional skills are malleable and responsive to pedagogical interventions (Sorrenti et al., 2020). Although the analysis centers on girls, parallel efforts that address how adolescent boys evaluate themselves and others are likely to be important for broader progress on gender equality.

Our results open several promising avenues for future research. First, longitudinal panel data could illuminate how these gendered evaluation dynamics evolve as students receive repeated feedback, both before and after engaging with personalized feedback reports. Such analysis would clarify the potential for dynamic feedback loops, a central feature in contemporary models of human capital formation (Coffman et al., 2024). Second, utilizing comparable data collected from control-group classrooms could establish whether these gender patterns are general features of adolescent development or specific responses to targeted classroom interventions. Lastly, exploring how demographic factors, such as student birthplace and parental education levels, influence peer assessments, and tracking how these relationships evolve over time, would further deepen our understanding of the nuanced interplay between gender, identity, and socioemotional skills.

Bibliography

References

Abrahams, L., Pancorbo, G., Primi, R., Santos, D., Kyllonen, P., John, O. P., and De Fruyt, F. (2019). Social-emotional skill assessment in children and adolescents: Advances and challenges in personality, clinical, and educational contexts. *Psychological assessment*, (4):460–473.

Ajayi, Kehinde, Das, Smita, Delavallade, Anne, C., Ketema, Assefa, T., Rouanet, and Marie, L. (2022). Gender differences in socio-emotional skills and economic outcomes : New evidence from 17 african countries (english). Technical report.

Alan, S. (2025). Shaping society's character: The role of schools in developing social and emotional skills. In B. Enke, P.Giuliano, N. Nunn and L. Wantchekon, editor, *Handbook of Culture*.

Andersen, S., Ertac, S., Gneezy, U., List, J., and Maximiano, S. (2013). Gender, competitiveness, and socialization at a young age: Evidence from a matrilineal and a patriarchal society. *The Review of Economics and Statistics*, (4):1438–1443.

Avitzour, E., Choen, A., Joel, D., and Lavy, V. (2020). On the origins of gender-biased behavior: The role of explicit and implicit stereotypes. *Social Science Research Network*.

Azmat, G., Iriberri, N., and Calsamiglia, C. (2016). Gender Differences in Response to Big Stakes. *Journal of the European Economic Association*, (6):1372–1400.

Babcock, L. and Laschever, S. (2021). Women don't ask: Negotiation and the gender divide.

Bagues, M., Sylos-Labini, M., and Zinovyeva, N. (2017). Does the gender composition of scientific committees matter? *American Economic Review*, (4):1207–1238.

Barber, B. M. and Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *The Quarterly Journal of Economics*, (1):261–292.

Benson, A., Li, D., and Shue, K. (2024). Potential and the gender promotions gap. *SSRN Electronic Journal*.

Beyer, S. and Bowden, E. M. (1997). Gender differences in seff-perceptions: Convergent evidence from three measures of accuracy and bias. *Personality & social psychology bulletin*, (2):157–172.

Blau, F. D. and Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, (3):789–865.

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of public economics*, pages 27–41.

Boring, A. and Ottoboni, K. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*.

Buser, T., van den Assem, M. J., and van Dolder, D. (2023). Gender and willingness to compete for high stakes. *Journal of Economic Behavior & Organization*, pages 350–370.

Cassidy, R., Das, S., Delavallade, C., Kipchumba, E., and Komba, J. (2024). Do men really have greater socio-emotional skills than women? evidence from tanzanian youth. Technical report.

Coffman, K., Ugalde Araya, M. P., and Zafar, B. (2024). A (dynamic) investigation of stereotypes, belief-updating, and behavior. *Economic inquiry*, (3):957–983.

Coffman, K. B. (2014). Evidence on self-stereotyping and the contribution of ideas. *The Quarterly Journal of Economics*, (4):1625–1660.

Cornwell, C., Mustard, D. B., and Van Parys, J. (2013). Noncognitive skills and the gender disparities in test scores and teacher assessments: Evidence from primary school. *The journal of human resources*, (1):236–264.

Delaney, J. M. and Devereux, P. J. (2021). Gender and educational achievement: Stylized facts and causal evidence.

Deming, D. J. (2017). The growing importance of social skills in the labor market. *The Quarterly Journal of Economics*, (4):1593–1640.

Dhar, D., Jain, T., and Jayachandran, S. (2022). Reshaping adolescents' gender attitudes: Evidence from a school-based experiment in india. *American Economic Review*, (3):899–927.

Duckworth, A. L. and Seligman, M. E. P. (2006). Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores. *Journal of educational psychology*, (1):198–208.

Eagly, A. H. and Wood, W. (2012). Social role theory. In *Handbook of Theories of Social Psychology*, pages 458–476. SAGE Publications Ltd, 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom.

Espey, M. (2022). Gender and peer evaluations. *The Journal of economic education*, (1):1–10.

Exley, C. L., Hauser, O. P., Moore, M., and Pezzuto, J.-H. (2024). Believed gender differences in social preferences. *The Quarterly Journal of Economics*.

Exley, C. L. and Kessler, J. B. (2022). The Gender Gap in Self-Promotion*. *The Quarterly Journal of Economics*, (3):1345–1381.

Falchikov, N. and Magin, D. (1997). Detecting gender bias in peer marking of students' group process work. *Assessment & Evaluation in Higher Education*, (4):385–396.

Fan, Y., Shepherd, L. J., Slavich, E., Waters, D., Stone, M., Abel, R., and Johnston, E. L. (2019). Gender and cultural bias in student evaluations: Why representation matters. *PloS one*, (2):e0209749.

Gneezy, U., Niederle, M., and Rustichini, A. (2003). Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics*, (3):1049–1074.

Goldin, C. (2014). A grand gender convergence: Its last chapter. *American Economic Review*, (4):1091–1119.

Goldin, C., Katz, L. F., and Kuziemko, I. (2006). The Homecoming of American College Women: The Reversal of the College Gender Gap. *The journal of economic perspectives: a journal of the American Economic Association*, (4):133–156.

Heckman, J. J. and Kautz, T. (2012). Hard evidence on soft skills. *Labour economics*.

Hernandez-Arenaz, I. and Iriberri, N. (2019). *A review of gender differences in negotiation*. Oxford University Press.

Hinnerich, B. T., Höglin, E., and Johannesson, M. (2011). Are boys discriminated in swedish high schools? *Economics of education review*, (4):682–690.

John, O., Naumann, L. P., and Soto, C. J. (2008). Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues. *Handbook of personality: Theory and research (3rd edition)*, pages 114–158.

Langan, A. M., Shuker, D. M., Cullen, W. R., Penney, D., Preziosi, R. F., and Wheater, C. P. (2008). Relationships between student characteristics and self-, peer and tutor evaluations of oral presentations. *Assessment and evaluation in higher education*, (2):179–190.

Langan, A. M., Wheater, C. P., Shaw, E. M., Haines, B. J., Cullen, W. R., Boyle, J. C., Penney, D., Oldekop, J. A., Ashcroft, C., Lockey, L., and Preziosi, R. F. (2005). Peer assessment of oral presentations: effects of student gender, university affiliation and participation in the development of assessment criteria. *Assessment and evaluation in higher education*, (1):21–34.

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? evidence from a natural experiment. *Journal of public economics*, (10-11):2083–2105.

MacNell, L., Driscoll, A., and Hunt, A. N. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative higher education*, (4):291–303.

Marianne, B. (2011). New perspectives on gender. In Card, D. and Ashenfelter, O., editors, *Handbook of Labor Economics*, Handbook of labour economics, pages 1543–1590. Elsevier.

Matavelli, I. (2025). We don't talk about boys: Masculinity norms among adolescents in brazil.

Mengel, F., Sauermann, J., and Zölitz, U. (2019). Gender Bias in Teaching Evaluations. *Journal of the European Economic Association*, (2):535–566.

Moreno, F. G. and Iñesta, E. L. (2021). Com classificar els centres educatius segons la complexitat? Technical report.

Murciano-Goroff, R. (2022). Missing women in tech: The labor market for highly skilled software engineers. *Management science*, (5):3262–3281.

Niederle, M. and Vesterlund, L. (2007). Do women shy away from competition? do men compete too much? *The Quarterly Journal of Economics*, (3):1067–1101.

Niederle, M. and Vesterlund, L. (2011). Gender and competition. *Annual Review of Economics*, pages 601–630.

OECD (2015). *Skills for social progress: the power of social and emotional skills*. OECD Skills Studies. Organization for Economic Co-operation and Development (OECD), Paris Cedex, France.

Ogden, T., Olseth, A., Sørlie, M.-A., and Hukkelberg, S. (2023). Teacher's assessment of gender differences in school performance, social skills, and externalizing behavior from fourth through seventh grade. *Journal of Education*, (1):211–221.

Porter, C. and Serra, D. (2020). Gender differences in the choice of major: The importance of female role models. *American Economic Journal. Applied Economics*, (3):226–254.

Recalde, M. P. and Vesterlund, L. (2023). Gender differences in negotiation: Can interventions reduce the gap? *Annual review of economics*, (1):633–657.

Reuben, E., Wiswall, M., and Zafar, B. (2017). Preferences and biases in educational choices and labour market expectations: Shrinking the black box of gender. *Economic journal (London, England)*, (604):2153–2186.

Roussille, N. (2024). The role of the ask gap in gender pay inequality. *The Quarterly Journal of Economics*, (3):1557–1610.

Rust, C., Price, M., and O'donovan, B. (2003). Improving students' learning by developing their understanding of assessment criteria and processes. *Assessment and evaluation in higher education*, (2):147–164.

Saygin, P. and Knight, T. (2023). Gender Bias in Performance Evaluations: Evidence from a Field Experiment. *Available at SSRN 4332175*.

Smith, B. and Wooten, J. (2024). Are Students Sexist when Rating Each Other? Bias in Peer Ratings and a Generalization of the Rubric-Based Estimator. *Social Science Research Network*.

Soll, J. B. and Klayman, J. (2004). Overconfidence in interval estimates. *Journal of experimental psychology. Learning, memory, and cognition*, (2):299–314.

Sorrenti, G., Zölitz, U., Ribeaud, D., and Eisner, M. (2020). The causal impact of socio-emotional skills training on educational success.

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*, (1):192–222.

Steinberg, L. (2014). *Age of opportunity: Lessons from the new science of adolescence*. Eamon Dolan/Houghton Mifflin Harcourt.

Torres-Guijarro, S. and Bengoechea, M. (2017). Gender differential in self-assessment: a fact neglected in higher education peer and self-assessment techniques. *Higher education research & development*, (5):1072–1084.

Tucker, R. (2014). Sex does not matter: gender bias and gender differences in peer assessments of contributions to group work. *Assessment and evaluation in higher education*, (3):293–309.

Wagner, N., Rieger, M., and Voorvelt, K. (2016). Gender, ethnicity and teaching evaluations: Evidence from mixed teaching teams. *Economics of education review*, pages 79–94.

Yeager, D. S. (2017). Social and emotional learning programs for adolescents. *The Future of children*, (1):73–94.

Appendix A

Standardization of socioemotional skills

The socioemotional skills dataset used in the analysis is a mixture of self-evaluations, peer-assigned scores, teacher assessments, and external observer ratings–all collected through the Pentabilities platform. In the classrooms, socioemotional skills assessments were collected across 35 behaviors (subdomains). To construct comparable and standardized measures, we aggregate these subdomains into the corresponding five socioemotional domains: Autonomy, Cooperation, Emotional Management, Thinking Abilities, and Responsibility. Below we provide the multi-step procedure to calculate the standardized peer-assigned score at the domain level, which serves as the main outcome variable in the analysis, and then also describe how these scores were calculated for each type of observer (i.e., self, teacher, and external observer).

$$b_{ijs} = \frac{1}{K}\sum_{k=1}^{K} b_{ijsk} \tag{A.1}$$

$$s_s = \frac{1}{N}\sum_{i=1}^{N} b_{ijs} \tag{A.2}$$

$$\sigma_s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(b_{ijs}-s_s)^2} \tag{A.3}$$

$$\widetilde{s_{ijs}} = \frac{(b_{ijs}-s_s)}{\sigma_s} \tag{A.4}$$

$$\widetilde{s_{ij}} = \frac{1}{S_{ij}}\sum_{s=1}^{S_{ij}} \widetilde{s_{ijs}} \tag{A.5}$$

$$\widetilde{\widetilde{s_{ij}}} = \frac{(s_{ij}-\widetilde{s})}{\sigma} \tag{A.6}$$

We denote b_{ijsk} as the k^{th} score of a subdomain s for student i assigned by peer j belonging to a particular socioemotional domain d (i.e., the raw score that ranges from 1 to 5). First, we compute the mean score assigned by each peer at the subdomain level, b_{ijs} (Equation A.1). Next, we standardize the subdomain-level scores within the sample. We compute the mean and standard deviation of subdomain scores across all students (Equations A.2 and A.3) and then construct a standardized score for each peer evaluation at the subdomain level using these moments, $\widetilde{s_{ijs}}$ (Equation A.4).

The standardized subdomain scores are then aggregated at the domain level for each student-peer pair, $\widetilde{s_{ij}}$, (Equation A.5). Finally, a second round of standardization is performed using the mean and standard deviation of the domain scores across all students $(\widetilde{s}, \sigma)$ to obtain the final measure as in Equation A.6. The resulting domain measure, $\widetilde{\widetilde{s_i}}$, is the standardized average of the standardized subdomain scores at the individual level.[30]

We use a similar two-step standardization procedure for the other types of evaluation. For self-assigned and teacher scores, first raw subdomain ratings from each evaluator were averaged for each student to produce a single subdomain-level score. This subdomain score was then standardized within the entire sample by subtracting the subdomain mean and dividing by the subdomain standard deviation. Next, the standardized subdomain scores were collapsed to the domain level by taking the mean of each student's standardized subdomain ratings and again standardized with respect to the domain-level mean and standard deviation across all students.

For the EO evaluations, we also repeat the same steps for the standardization, but this time we use as moments the subdomain mean and standard deviation in the control group (s_s^C, σ_s^C) and we also standardize the domain score using the moments in the control group $(\widetilde{s^C}, \sigma^C)$. The final score measures thus mirror the construction of the peer-assigned scores, ensuring all four evaluator types are placed on comparable scales.

10-item survey

After the standardized activity, students are asked to complete a survey rating from 1 to 5 their performance during the activity on 10 behaviors. They are also asked to rate the behaviors of all their peers in their small group. The external observer that evaluated the behaviors of that small group during the standardized activity also completes the same summative survey for the same students. Students and observers are allowed to not answer or select "I don't know / I can't rate

[30] Note that we do not impute values to missing subdomain observations. Hence, the average domain score is computed among the number of subdomains scored (s_i).

that". External observers also answer an additional question for each student that assesses whether their answer is reliable given the behavior (or misbehavior) during the survey completion.

The objective of the summative survey is to provide a summary snapshot of the behaviors of the students from the perspective of self, peer, and external observer. We attempted to design a brief survey—-with only 10-items—that captured the five Pentabilities behavioral domains, without explicitly referring to them,[31] and taking the wording from validated socioemotional skill surveys. Hence, the 10 behaviors of the summative survey are a subset of the behaviors in the BESSI survey (192 item), which have been selected through an iterative mapping and triangulation process between the 35 Pentabilities behaviors, the BESSI Survey (192 item) and the behaviors that can be feasibly elicited in the standardized activity given its design. The selection process of the 10 BESSI behaviors to be used for the summative survey has been validated by an expert in education, psychopedagogy and emotional intelligence in education. In order to aid the understanding of the behavior wording, the expert proposed extreme single-sentence examples for each behavior, one for a rating of 1 and another for a rating of 5. We use the data from the summative surveys (from self, peers, and external observers) to construct measures of individuals' awareness of one's skills.

The 10 items in the summative survey are the following:

- Kept myself from getting distracted
  - Example for 1: I got distracted and wasted my time.
  - Example for 5: I was focused and made the most of my time

- Worked hard to succeed
  - Example for 1: I didn't put any effort into the task
  - Example for 5: I put a lot of effort carrying out the task.

- Worked with people towards a shared goal
  - Example for 1: We did not agree and have not been able to work well together.
  - Example for 5: We have cooperated and worked very well together.

[31] Since the original study is an RCT, this was designed so that the treatment students who used Pentabilities app were not given an undue advantage.

- Stayed calm in stressful situations
    - Example for 1: I became very agitated in stressful situations.
    - Example for 5: I remained calm in stressful stuations.
- Explained what I am thinking and feeling
    - Example for 1: I did not express what I felt and thought at any moment.
    - Example for 5: I did express what I felt and thought at all times
- Followed the rules
    - Example for 1: In many occasions I did not follow the rules
    - Example for 5: I followed the rules at all times.
- Changed people's mind
    - Example for 1: My proposals were never accepted by my team.
    - Example for 5: My proposals were always accepted by my team.
- Have taken another person's perspective
    - Example for 1: I did not see or consider how others felt at any point.
    - Example for 5: I have seen and considered how others feel at all times.
- Took responsibility when I made a mistake
    - Example for 1: When I made a mistake I always made excuses or blamed others and did not try to fix it.
    - Example for 5: When I made a mistake I always accepted it and tried to fix it.
- Found logical solutions to problems
    - Example for 1: I was not able to find good solutions to problems.
    - Example for 5: I was able to find good solutions to problems.

Appendix B

Descriptive Statistics: Standardized Pentabilities score across 4 raters

Table B.14: Summary Statistics: Self-Evaluation by Gender

	Female			Male			Total				
	Mean	**SD**	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**	**Min**	**Max**	**N**
Autonomy	-.02	.98	302	.01	1	338	0	1	-3.79	1.23	673
Cooperation	0	1.02	364	-.01	.96	388	0	1	-4.41	1.27	783
Responsibility	.03	.99	392	-.02	1	428	0	1	-4.44	1.25	861
Emt. Mngt.	-.13	1.04	292	.09	.96	326	0	1	-3.75	1.26	648
Thk. Abi.	-.08	1.01	303	.08	.97	304	0	1	-3.79	1.43	635

Note: The total scores reflect standardization at the domain level so that the pooled mean is $\widetilde{s} = 0$ and the standard deviation is $\sigma = 1$.

Table B.15: Unequal variance t-test (Female vs. Male): Self-Evaluation

Variable	Diff	SE	p-value
Autonomy	-0.03	0.08	0.695
Cooperation	0.01	0.07	0.904
Responsibility	0.05	0.07	0.453
Emt. Mngt.	-0.21***	0.08	0.008
Thk. Abi.	-0.17**	0.08	0.039

Table B.16: Summary Statistics: Peer Evaluations

	Female Students			Male Students			Total				
	Mean	**SD**	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**	**Min**	**Max**	**N**
Peer Evalrs/ Student	7.24	5.52	472	6.93	5.33	513	7.04	5.38	1.00	27.00	1031
Panel A: All Peers											
Autonomy	0.12	0.94	1917	-0.12	1.04	1923	-0.00	1.00	-2.44	1.05	3999
Cooperation	0.14	0.94	2609	-0.12	1.03	2663	0.00	1.00	-2.70	1.09	5511
Responsibility	0.17	0.92	2659	-0.15	1.04	2737	-0.00	1.00	-2.79	1.05	5619
Emt. Mngt.	0.07	0.97	2309	-0.07	1.02	2434	0.00	1.00	-2.55	1.14	4951
Thk. Abi.	0.09	0.96	2090	-0.08	1.01	2070	-0.00	1.00	-2.57	1.17	4288
Panel B: Female Peers											
Autonomy	0.25	0.84	850	-0.20	1.06	637	0.05	0.97	-2.44	1.05	1543
Cooperation	0.25	0.88	1086	-0.22	1.03	938	0.02	0.99	-2.70	1.09	2110
Responsibility	0.21	0.91	1077	-0.26	1.04	935	-0.01	1.00	-2.79	1.05	2085
Emt. Mngt.	0.07	0.97	967	-0.22	1.01	890	-0.06	1.00	-2.55	1.14	1929
Thk. Abi.	0.09	0.95	954	-0.29	1.00	779	-0.08	0.99	-2.57	1.17	1775
Panel C: Male Peers											
Autonomy	0.04	0.99	640	-0.01	1.00	859	0.00	1.00	-2.44	1.05	1556
Cooperation	0.09	0.97	897	-0.02	1.00	1110	0.02	0.99	-2.70	1.09	2098
Responsibility	0.14	0.94	910	-0.04	1.02	1122	0.03	0.99	-2.79	1.05	2120
Emt. Mngt.	0.03	0.94	812	0.05	0.99	1009	0.03	0.97	-2.55	1.14	1900
Thk. Abi.	0.04	0.97	686	0.07	0.96	872	0.05	0.98	-2.57	1.17	1605

Table B.17: Unequal variance t-tests (Female vs. Male Students): Peer Evaluations

	Diff	SE	p-value
Panel A: All Peers			
Autonomy	0.24***	0.03	0.000
Cooperation	0.26***	0.03	0.000
Responsibility	0.32***	0.03	0.000
Emt. Mngt.	0.14***	0.03	0.000
Thk. Abi.	0.17***	0.03	0.000
Panel B: Female Peers			
Autonomy	0.45***	0.05	0.000
Cooperation	0.46***	0.04	0.000
Responsibility	0.47***	0.04	0.000
Emt. Mngt.	0.29***	0.05	0.000
Thk. Abi.	0.38***	0.05	0.000
Panel C: Male Peers			
Autonomy	0.05	0.05	0.352
Cooperation	0.11**	0.04	0.014
Responsibility	0.18***	0.04	0.000
Emt. Mngt.	-0.02	0.05	0.617
Thk. Abi.	-0.03	0.05	0.510

Table B.18: Summary Statistics: Teacher Evaluation

	Female Student			Male Student			Total				
	Mean	**SD**	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**	**Min**	**Max**	**N**
Tch Evalrs/ Student	1.54	0.79	519	1.62	0.86	568	1.58	0.83	1.00	6.00	1136
Panel A: All Teachers											
Autonomy	0.17	0.98	554	-0.15	0.99	611	0.00	1.00	-1.94	1.32	1210
Cooperation	0.25	0.91	632	-0.21	1.01	700	0.00	1.00	-2.19	1.36	1395
Responsibility	0.27	0.89	736	-0.22	1.02	844	0.00	1.00	-2.33	1.25	1651
Emt. Mngt.	0.16	0.96	475	-0.13	1.00	532	0.00	1.00	-2.24	1.37	1059
Thk. Abi.	0.22	0.95	462	-0.20	1.00	495	0.00	1.00	-1.87	1.45	990
Panel B: Female Teachers											
Autonomy	0.27	0.94	321	-0.15	1.01	296	0.06	0.99	-1.94	1.32	645
Cooperation	0.31	0.86	359	-0.25	1.03	364	0.02	0.99	-2.19	1.36	754
Responsibility	0.29	0.85	403	-0.24	1.04	405	0.01	1.00	-2.33	1.25	844
Emt. Mngt.	0.17	0.97	277	-0.16	0.97	279	0.01	0.99	-2.24	1.37	579
Thk. Abi.	0.24	0.95	282	-0.18	1.01	263	0.03	1.00	-1.87	1.45	567
Panel C: Male Teachers											
Autonomy	-0.19	1.09	104	-0.16	1.00	159	-0.16	1.03	-1.94	1.32	271
Cooperation	-0.01	1.03	97	-0.04	0.92	131	-0.04	0.97	-2.19	1.36	241
Responsibility	0.12	0.96	149	-0.13	0.94	210	-0.04	0.97	-2.33	1.25	374
Emt. Mngt.	0.04	0.93	69	0.15	0.87	108	0.06	0.91	-2.24	1.37	191
Thk. Abi.	0.01	1.07	65	0.00	0.97	93	0.00	1.01	-1.87	1.45	160

Table B.19: Welch Unequal-Variance t-tests (Female vs. Male Students): Teacher Evaluations

	Diff (F–M)	SE	p-value
Panel A: All Teachers			
Autonomy	0.32***	0.06	0.000
Cooperation	0.46***	0.05	0.000
Responsibility	0.49***	0.05	0.000
Emt. Mngt.	0.29***	0.06	0.000
Thk. Abi.	0.42***	0.06	0.000
Panel B: Female Teachers			
Autonomy	0.41***	0.08	0.000
Cooperation	0.56***	0.07	0.000
Responsibility	0.53***	0.07	0.000
Emt. Mngt.	0.32***	0.08	0.000
Thk. Abi.	0.41***	0.08	0.000
Panel C: Male Teachers			
Autonomy	-0.03	0.13	0.847
Cooperation	0.03	0.13	0.793
Responsibility	0.25**	0.10	0.015
Emt. Mngt.	-0.10	0.14	0.456
Thk. Abi.	0.01	0.17	0.945

Table B.20: Summary Statistics for EO Scores by Gender

	Female Students			**Male Students**			**Total**				
	Mean	**SD**	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**	**Min**	**Max**	**N**
EOs/Student	2.22	1.18	303	2.25	1.18	304	2.22	1.17	1.00	6.00	631
Panel A: All EO											
Autonomy	0.07	0.98	550	-0.07	1.00	564	-0.01	1.00	-1.88	1.93	1144
Cooperation	0.13	0.97	526	-0.12	0.99	520	-0.01	1.00	-2.18	1.86	1079
Responsibility	0.15	0.97	671	-0.15	1.00	681	0.00	1.00	-2.71	1.85	1393
Emotional Management	0.10	0.98	395	-0.08	1.01	436	0.00	1.00	-2.22	2.05	851
Thinking Abilities	0.08	0.95	332	-0.06	1.04	338	0.00	1.00	-2.05	2.02	687
Panel B: Female EO											
Autonomy	0.08	0.97	408	0.00	1.04	405	0.03	1.01	-1.88	1.93	833
Cooperation	0.10	0.99	394	-0.10	1.04	369	-0.01	1.02	-2.18	1.86	785
Responsibility	0.09	0.98	516	-0.16	1.02	510	-0.04	1.01	-2.71	1.85	1057
Emotional Management	0.13	0.99	275	-0.06	1.06	292	0.03	1.03	-2.22	2.05	579
Thinking Abilities	0.14	0.93	251	0.10	1.05	237	0.11	0.99	-2.05	2.02	498
Panel C: Male EO											
Autonomy	0.03	1.03	142	-0.24	0.86	159	-0.12	0.96	-1.88	1.93	311
Cooperation	0.22	0.93	132	-0.18	0.86	151	-0.01	0.92	-2.18	1.86	294
Responsibility	0.32	0.93	155	-0.11	0.93	171	0.10	0.96	-2.71	1.85	336
Emotional Management	0.02	0.96	120	-0.10	0.91	144	-0.06	0.93	-2.22	2.05	272
Thinking Abilities	-0.11	1.00	81	-0.42	0.93	101	-0.27	0.96	-2.05	2.02	189

Table B.21: Welch Unequal-Variance t-tests (Female vs. Male) for EO Evaluations

	Diff (F–M)	SE	p-value
Panel A: All EO			
Autonomy	0.14**	0.06	0.020
Cooperation	0.25***	0.06	0.000
Responsibility	0.29***	0.05	0.000
Emotional Management	0.17**	0.07	0.012
Thinking Abilities	0.13*	0.08	0.082
Panel B: Female EO			
Autonomy	0.09	0.07	0.219
Cooperation	0.20***	0.07	0.007
Responsibility	0.25***	0.06	0.000
Emotional Management	0.19**	0.09	0.024
Thinking Abilities	0.04	0.09	0.642
Panel C: Male EO			
Autonomy	0.27**	0.11	0.016
Cooperation	0.40***	0.11	0.000
Responsibility	0.43***	0.10	0.000
Emotional Management	0.13	0.12	0.279
Thinking Abilities	0.31**	0.15	0.037

CDFs for self and peer score distributions

The CDFs for self-assigned score depicted in Figure B.1. In the two domains that we observe gender differences in scores, i.e., Emotional management and Thinking abilities, males self-assign slightly better scores at higher levels, whereas females appear more concentrated around lower-to-average scores.

Figure B.1: CDFs of Self-assigned Score

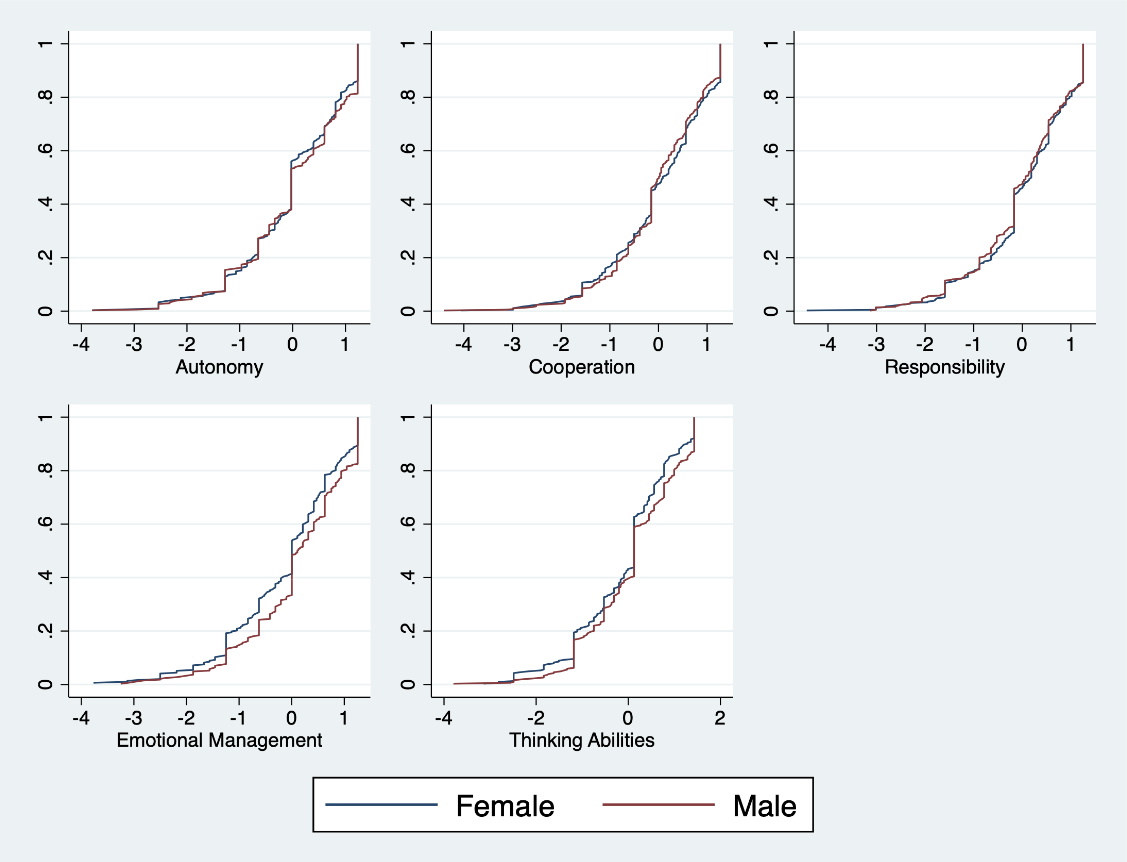


Figure B.2 displays peer-assigned scores by student gender. Female students receive higher scores across the distribution, with their CDF stochastically dominating that of male students in all domains. Figure B.3 further breaks down peer-assigned scores by gender of the peer and the student, revealing that the largest divergences emerge in the mid-to-upper range of the distribution, particularly for scores assigned by female peers to male vs female students.

Figure B.2: CDFs of Peer-assigned Score: Students

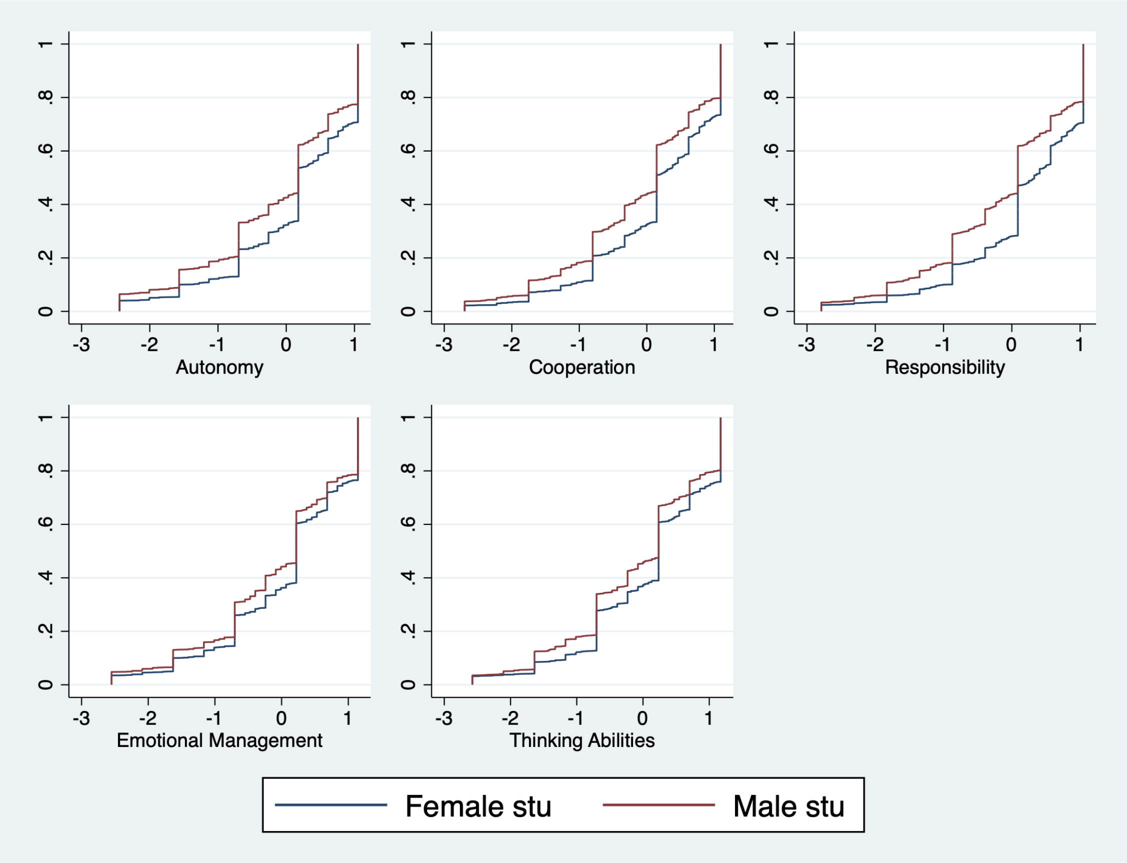
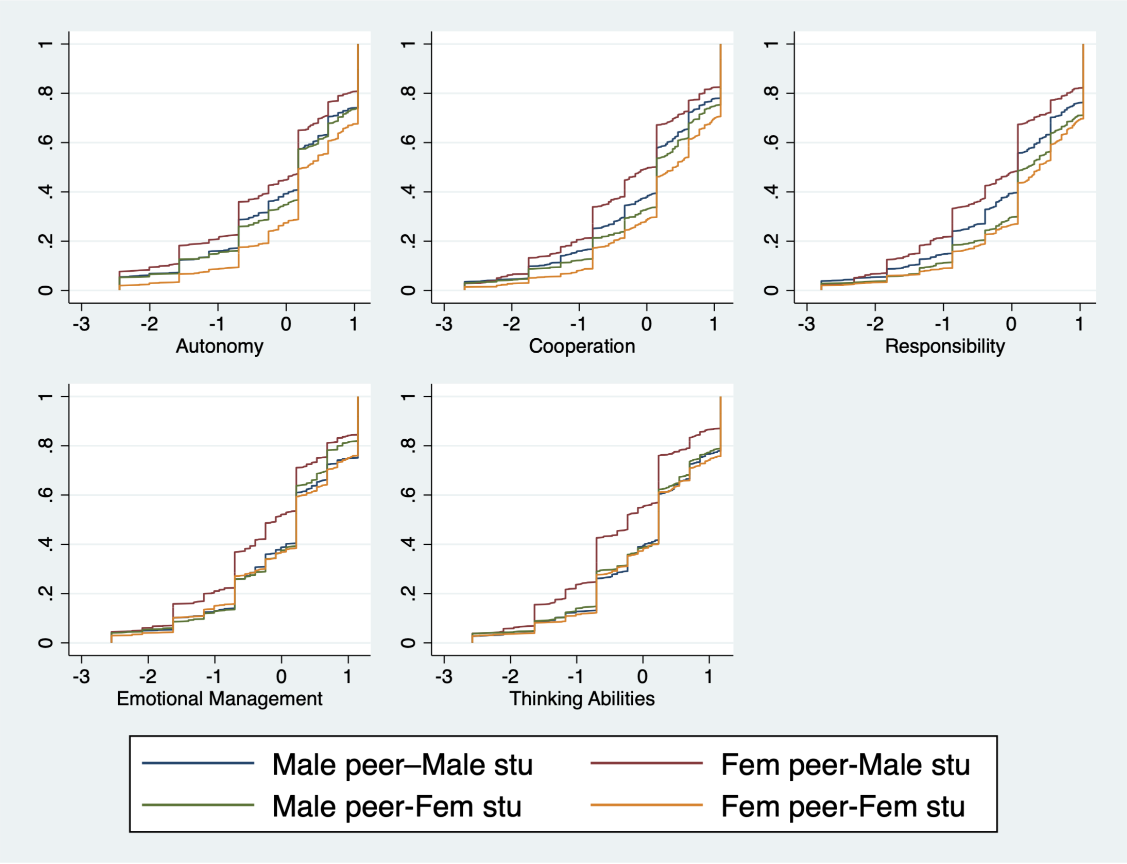


Figure B.3: CDFs of Peer-assigned Score: Gender Pairs



Appendix C

Full Regression Tables

Self-assigned score

Table C.22: Self-Assigned Score - Autonomy

	(1)	(2)	(3)
Female Student	−0.058	−0.141*	−0.141*
	(0.081)	(0.080)	(0.081)
Student's Teacher Score		0.301***	0.177***
		(0.053)	(0.061)
Ave Peer Score: autonomy			0.270***
			(0.080)
Student born in Spain	0.148	0.080	0.040
	(0.157)	(0.158)	(0.157)
Student Caregiver 1: Born in Spain	0.259**	0.248*	0.218*
	(0.130)	(0.132)	(0.132)
Student Caregiver 2: Born in Spain	0.033	−0.013	0.015
	(0.139)	(0.141)	(0.140)
Stu CG1 went to uni	0.130	0.083	0.091
	(0.101)	(0.103)	(0.105)
Stu CG1 uni edu not known	0.161	0.131	0.118
	(0.124)	(0.123)	(0.122)
Stu CG2 went to uni	0.190*	0.140	0.153
	(0.101)	(0.102)	(0.103)
Stu CG2 uni edu not known	−0.218**	−0.236**	−0.164
	(0.108)	(0.109)	(0.111)
Constant	−0.372**	−0.189	−0.182
	(0.147)	(0.152)	(0.155)
Classroom FE	*Yes*	*Yes*	*Yes*
R-squared	0.213	0.266	0.286
Observations	570	540	525

Note: Robust standard errors in parentheses. * (p<0.10), ** (p<0.05), *** (p<0.01).

Table C.23: Self-Assigned Score - Cooperation

	(1)	(2)	(3)
Female Student	0.011	−0.075	−0.101
	(0.073)	(0.079)	(0.078)
Student's Teacher Score		0.217***	0.117**
		(0.054)	(0.058)
Ave Peer Score: cooperation			0.290***
			(0.074)
Student born in Spain	0.228	0.202	0.183
	(0.140)	(0.141)	(0.144)
Student Caregiver 1: Born in Spain	0.172	0.120	0.098
	(0.126)	(0.134)	(0.132)
Student Caregiver 2: Born in Spain	−0.108	−0.085	−0.088
	(0.126)	(0.133)	(0.130)
Stu CG1 went to uni	0.134	0.130	0.126
	(0.089)	(0.088)	(0.087)
Stu CG1 uni edu not known	0.070	0.083	0.060
	(0.110)	(0.108)	(0.106)
Stu CG2 went to uni	0.082	0.041	0.066
	(0.090)	(0.089)	(0.088)
Stu CG2 uni edu not known	−0.163	−0.158	−0.090
	(0.105)	(0.103)	(0.102)
Constant	−0.275*	−0.182	−0.158
	(0.146)	(0.148)	(0.153)
Classroom FE	*Yes*	*Yes*	*Yes*
R-squared	0.241	0.263	0.287
Observations	668	646	628

Note: Robust standard errors in parentheses. * (p<0.10), ** (p<0.05), *** (p<0.01).

Table C.24: Self-Assigned Score - Responsibility

	(1)	(2)	(3)
Female Student	0.080	−0.085	−0.085
	(0.072)	(0.075)	(0.074)
Student's Teacher Score		0.311***	0.137***
		(0.047)	(0.053)
Ave Peer Score: responsibility			0.384***
			(0.068)
Student born in Spain	0.307**	0.255**	0.200
	(0.132)	(0.129)	(0.131)
Student Caregiver 1: Born in Spain	0.130	0.074	0.069
	(0.133)	(0.133)	(0.126)
Student Caregiver 2: Born in Spain	−0.057	−0.076	−0.046
	(0.135)	(0.132)	(0.127)
Stu CG1 went to uni	0.080	0.079	0.068
	(0.091)	(0.091)	(0.090)
Stu CG1 uni edu not known	0.002	−0.000	0.058
	(0.121)	(0.117)	(0.118)
Stu CG2 went to uni	0.122	0.061	0.029
	(0.094)	(0.093)	(0.093)
Stu CG2 uni edu not known	−0.069	−0.054	−0.064
	(0.112)	(0.109)	(0.109)
Constant	−0.414***	−0.247*	−0.211
	(0.135)	(0.131)	(0.133)
Classroom FE	*Yes*	*Yes*	*Yes*
R-squared	0.201	0.257	0.305
Observations	729	714	692

Note: Robust standard errors in parentheses. * (p<0.10), ** (p<0.05), *** (p<0.01).

Table C.25: Self-Assigned Score - Emotional Management

	(1)	(2)	(3)
Female Student	-0.217^{**}	-0.290^{***}	-0.302^{***}
	(0.085)	(0.091)	(0.090)
Student’s Teacher Score		0.200^{***}	0.087
		(0.058)	(0.058)
Ave Peer Score: emotion			0.319^{***}
			(0.089)
Student born in Spain	0.072	0.017	-0.017
	(0.154)	(0.156)	(0.158)
Student Caregiver 1: Born in Spain	-0.051	-0.053	-0.094
	(0.151)	(0.162)	(0.169)
Student Caregiver 2: Born in Spain	0.093	0.139	0.176
	(0.150)	(0.158)	(0.164)
Stu CG1 went to uni	0.016	-0.042	-0.041
	(0.100)	(0.103)	(0.101)
Stu CG1 uni edu not known	0.034	0.092	0.144
	(0.132)	(0.133)	(0.131)
Stu CG2 went to uni	0.229^{**}	0.287^{**}	0.274^{**}
	(0.107)	(0.113)	(0.113)
Stu CG2 uni edu not known	-0.078	-0.044	-0.017
	(0.126)	(0.127)	(0.121)
Constant	-0.078	-0.056	-0.033
	(0.167)	(0.175)	(0.174)
Classroom FE	*Yes*	*Yes*	*Yes*
R-squared	0.224	0.249	0.280
Observations	551	481	462

Note: Robust standard errors in parentheses. * (p<0.10), ** (p<0.05), *** (p<0.01).

Table C.26: Self-Assigned Score - Thinking Abilities

	(1)	(2)	(3)
Female Student	-0.148^{*}	-0.204^{**}	-0.239^{***}
	(0.082)	(0.091)	(0.090)
Student's Teacher Score		0.175^{***}	0.042
		(0.059)	(0.066)
Ave Peer Score: thinking			0.306^{***}
			(0.080)
Student born in Spain	-0.065	-0.145	-0.199
	(0.154)	(0.157)	(0.154)
Student Caregiver 1: Born in Spain	0.285^{*}	0.315^{*}	0.235
	(0.151)	(0.166)	(0.164)
Student Caregiver 2: Born in Spain	-0.207	-0.244	-0.164
	(0.152)	(0.170)	(0.164)
Stu CG1 went to uni	0.023	0.018	0.047
	(0.099)	(0.103)	(0.102)
Stu CG1 uni edu not known	-0.008	-0.048	-0.023
	(0.122)	(0.132)	(0.128)
Stu CG2 went to uni	0.177^{*}	0.140	0.139
	(0.100)	(0.103)	(0.101)
Stu CG2 uni edu not known	-0.158	-0.133	-0.113
	(0.110)	(0.117)	(0.112)
Constant	0.058	0.158	0.214
	(0.161)	(0.166)	(0.162)
Classroom FE	*Yes*	*Yes*	*Yes*
R-squared	0.286	0.313	0.329
Observations	544	479	468

Note: Robust standard errors in parentheses. * (p<0.10), ** (p<0.05), *** (p<0.01).

Table C.27: Self-Assigned Score: Teacher Gender × Teacher Score (5 domains)

	Autonomy	Cooperation	Responsibility	Emotion	Thinking
Female Student	-0.124	-0.088	-0.071	-0.253***	-0.223**
	(0.085)	(0.084)	(0.080)	(0.095)	(0.096)
Female Teacher	-0.207	0.139	-0.002	-0.286	0.185
	(0.187)	(0.158)	(0.178)	(0.177)	(0.233)
Teacher Score	0.056	0.129	0.166*	0.062	0.014
	(0.107)	(0.115)	(0.093)	(0.112)	(0.199)
Teacher Female × Teacher Score	0.133	0.003	-0.018	-0.025	0.017
	(0.118)	(0.129)	(0.103)	(0.127)	(0.201)
Average Peer Score	0.307***	0.332***	0.394***	0.379***	0.347***
	(0.079)	(0.081)	(0.072)	(0.095)	(0.086)
Observations	493	583	647	434	423
Adj. R-squared	0.193	0.143	0.208	0.153	0.203

Notes: All regressions control for student background characteristics and school fixed effects. Robust standard errors in parentheses. The dependent variable is self-assigned score. *p<0.10, ** p<0.05, *** p<0.010.

Peer-assigned score

Table C.28: Peer-Assigned Score - Autonomy

	(1)	(2)	(3)	(4)	(5)
Female Student (β_1)	0.034	0.050	−0.082	−0.088	−0.086
	(0.069)	(0.072)	(0.065)	(0.065)	(0.065)
Female Peer (β_2)	−0.212***	−0.210***	−0.225***	−0.235***	−0.207***
	(0.065)	(0.067)	(0.063)	(0.064)	(0.064)
Female Student x Female Peer (β_3)	0.438***	0.423***	0.416***	0.446***	0.454***
	(0.084)	(0.086)	(0.082)	(0.083)	(0.084)
Peer's Self Score		0.145***	0.154***	0.167***	0.183***
		(0.026)	(0.025)	(0.025)	(0.026)
Student's Teacher Score			0.416***	0.361***	0.364***
			(0.034)	(0.038)	(0.038)
Student's Self Score				0.121***	0.116***
				(0.040)	(0.040)
Peer's Teacher Score					−0.083***
					(0.028)
Student born in Spain	0.198**	0.185**	0.123	0.072	0.100
	(0.090)	(0.093)	(0.077)	(0.081)	(0.081)
Student Caregiver 1: Born in Spain	0.202**	0.187*	0.103	0.111	0.110
	(0.094)	(0.100)	(0.086)	(0.084)	(0.085)
Student Caregiver 2: Born in Spain	0.008	0.049	0.017	−0.002	−0.008
	(0.098)	(0.104)	(0.088)	(0.090)	(0.091)
Stu CG1 went to uni	0.035	0.039	0.007	0.031	0.036
	(0.069)	(0.071)	(0.067)	(0.067)	(0.068)
Stu CG1 uni edu not known	−0.146*	−0.167*	−0.176**	−0.152*	−0.145*
	(0.087)	(0.090)	(0.083)	(0.086)	(0.087)
Stu CG2 went to uni	0.147**	0.161**	0.115*	0.113*	0.122*
	(0.070)	(0.072)	(0.066)	(0.067)	(0.068)
Stu CG2 uni edu not known	0.018	0.046	0.045	0.080	0.085
	(0.085)	(0.087)	(0.077)	(0.080)	(0.081)
Peer born in Spain	−0.089	−0.081	−0.108*	−0.147**	−0.158**
	(0.061)	(0.063)	(0.062)	(0.063)	(0.064)
Peer Caregiver 1: Born in Spain	0.110	0.120	0.127	0.153*	0.167**
	(0.079)	(0.081)	(0.079)	(0.081)	(0.080)
Peer Caregiver 2: Born in Spain	−0.055	−0.089	−0.122	−0.119	−0.097
	(0.081)	(0.082)	(0.079)	(0.080)	(0.080)
Peer CG1 went to uni	0.107**	0.105**	0.104**	0.093*	0.086*
	(0.047)	(0.050)	(0.047)	(0.048)	(0.049)
Peer CG1 uni edu not known	0.024	−0.007	0.000	−0.011	−0.016
	(0.057)	(0.060)	(0.056)	(0.058)	(0.058)
Peer CG2 went to uni	−0.010	−0.036	−0.048	−0.064	−0.037
	(0.051)	(0.053)	(0.051)	(0.051)	(0.052)
Peer CG2 uni edu not known	−0.044	−0.021	−0.023	−0.033	−0.020
	(0.057)	(0.060)	(0.057)	(0.059)	(0.060)
Constant	−0.326**	−0.332**	−0.113	−0.056	−0.113
	(0.132)	(0.134)	(0.116)	(0.120)	(0.122)
Fem Peer: Fem vs Male Stu ($\beta_1 + \beta_3$)	0.472	0.472	0.334	0.358	0.368
Fem Peer: P-val	0.000	0.000	0.000	0.000	0.000
Classroom FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
R-squared	0.170	0.191	0.283	0.281	0.286
Observations	2420	2232	2184	2038	2006

Standard errors in parentheses and clustered at student level. * (p<0.10), ** (p<0.05), *** (p<0.01).

Table C.29: Peer-Assigned Score - Cooperation

	(1)	(2)	(3)	(4)	(5)
Female Student (β_1)	0.092	0.062	-0.117^{**}	-0.097^{*}	-0.112^{**}
	(0.057)	(0.058)	(0.054)	(0.055)	(0.056)
Female Peer (β_2)	-0.228^{***}	-0.242^{***}	-0.257^{***}	-0.273^{***}	-0.293^{***}
	(0.055)	(0.053)	(0.050)	(0.051)	(0.053)
Female Student x Female Peer (β_3)	0.338^{***}	0.362^{***}	0.376^{***}	0.390^{***}	0.406^{***}
	(0.070)	(0.069)	(0.066)	(0.067)	(0.068)
Peer's Self Score		0.266^{***}	0.272^{***}	0.273^{***}	0.264^{***}
		(0.021)	(0.019)	(0.020)	(0.021)
Student's Teacher Score			0.401^{***}	0.351^{***}	0.349^{***}
			(0.033)	(0.033)	(0.033)
Student's Self Score				0.088^{***}	0.091^{***}
				(0.029)	(0.029)
Peer's Teacher Score					0.019
					(0.027)
Student born in Spain	0.227^{***}	0.222^{**}	0.247^{***}	0.235^{***}	0.242^{***}
	(0.085)	(0.086)	(0.077)	(0.080)	(0.081)
Student Caregiver 1: Born in Spain	0.230^{***}	0.251^{***}	0.160^{**}	0.194^{***}	0.206^{***}
	(0.081)	(0.084)	(0.077)	(0.071)	(0.070)
Student Caregiver 2: Born in Spain	-0.043	-0.043	-0.065	-0.093	-0.113
	(0.081)	(0.083)	(0.076)	(0.073)	(0.071)
Stu CG1 went to uni	-0.068	-0.077	-0.065	-0.076	-0.087
	(0.061)	(0.062)	(0.054)	(0.055)	(0.055)
Stu CG1 uni edu not known	-0.190^{***}	-0.221^{***}	-0.162^{***}	-0.194^{***}	-0.196^{***}
	(0.070)	(0.070)	(0.062)	(0.062)	(0.062)
Stu CG2 went to uni	0.157^{**}	0.151^{**}	0.086	0.090	0.076
	(0.063)	(0.065)	(0.057)	(0.057)	(0.057)
Stu CG2 uni edu not known	0.020	0.044	0.067	0.082	0.075
	(0.068)	(0.069)	(0.062)	(0.063)	(0.063)
Peer born in Spain	-0.153^{***}	-0.153^{***}	-0.175^{***}	-0.171^{***}	-0.164^{***}
	(0.056)	(0.052)	(0.052)	(0.054)	(0.055)
Peer Caregiver 1: Born in Spain	0.041	0.034	0.034	0.018	0.019
	(0.058)	(0.058)	(0.059)	(0.060)	(0.061)
Peer Caregiver 2: Born in Spain	0.103^{*}	0.098^{*}	0.096^{*}	0.111^{*}	0.109^{*}
	(0.059)	(0.057)	(0.056)	(0.058)	(0.058)
Peer CG1 went to uni	0.066	0.079^{*}	0.062	0.055	0.063
	(0.041)	(0.042)	(0.042)	(0.042)	(0.042)
Peer CG1 uni edu not known	-0.020	-0.024	-0.019	-0.023	-0.023
	(0.050)	(0.049)	(0.049)	(0.050)	(0.051)
Peer CG2 went to uni	0.068^{*}	0.029	-0.000	-0.001	-0.016
	(0.040)	(0.040)	(0.039)	(0.039)	(0.039)
Peer CG2 uni edu not known	-0.007	0.044	0.012	0.009	0.006
	(0.044)	(0.044)	(0.043)	(0.044)	(0.045)
Constant	-0.286^{**}	-0.298^{**}	-0.163	-0.147	-0.121
	(0.124)	(0.127)	(0.113)	(0.118)	(0.119)
Fem Peer: Fem vs Male Stu ($\beta_1 + \beta_3$)	0.430	0.424	0.259	0.293	0.294
Fem Peer: P-val	0.000	0.000	0.000	0.000	0.000
Classroom FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
R-squared	0.146	0.195	0.267	0.267	0.267
Observations	3297	3111	3040	2891	2842

Standard errors in parentheses and clustered at student level. * (p<0.10), ** (p<0.05), *** (p<0.01).

Table C.30: Peer-Assigned Score - Responsibility

	(1)	(2)	(3)	(4)	(5)
Female Student (β_1)	0.227***	0.223***	−0.047	−0.025	−0.019
	(0.060)	(0.061)	(0.056)	(0.056)	(0.056)
Female Peer (β_2)	−0.197***	−0.208***	−0.195***	−0.201***	−0.156***
	(0.056)	(0.056)	(0.051)	(0.052)	(0.055)
Female Student x Female Peer (β_3)	0.261***	0.244***	0.242***	0.246***	0.232***
	(0.074)	(0.074)	(0.068)	(0.069)	(0.070)
Peer's Self Score		0.201***	0.195***	0.192***	0.205***
		(0.022)	(0.020)	(0.021)	(0.021)
Student's Teacher Score			0.466***	0.414***	0.418***
			(0.031)	(0.030)	(0.030)
Student's Self Score				0.095***	0.094***
				(0.030)	(0.030)
Peer's Teacher Score					−0.061**
					(0.025)
Student born in Spain	0.281***	0.297***	0.160**	0.105	0.111
	(0.091)	(0.092)	(0.073)	(0.079)	(0.078)
Student Caregiver 1: Born in Spain	0.114	0.109	0.013	0.034	0.032
	(0.083)	(0.085)	(0.074)	(0.073)	(0.071)
Student Caregiver 2: Born in Spain	0.010	0.020	0.036	0.025	0.030
	(0.081)	(0.082)	(0.071)	(0.070)	(0.068)
Stu CG1 went to uni	−0.039	−0.078	−0.043	−0.051	−0.053
	(0.070)	(0.070)	(0.060)	(0.061)	(0.061)
Stu CG1 uni edu not known	−0.192**	−0.235***	−0.177**	−0.149**	−0.151**
	(0.081)	(0.083)	(0.069)	(0.072)	(0.072)
Stu CG2 went to uni	0.209***	0.210***	0.118**	0.113*	0.112*
	(0.071)	(0.072)	(0.059)	(0.060)	(0.060)
Stu CG2 uni edu not known	0.085	0.104	0.108	0.078	0.077
	(0.080)	(0.082)	(0.071)	(0.072)	(0.072)
Peer born in Spain	−0.045	−0.064	−0.085	−0.102*	−0.087
	(0.054)	(0.053)	(0.053)	(0.054)	(0.055)
Peer Caregiver 1: Born in Spain	0.087	0.071	0.073	0.078	0.081
	(0.059)	(0.059)	(0.057)	(0.058)	(0.059)
Peer Caregiver 2: Born in Spain	0.008	0.001	0.002	0.003	0.007
	(0.061)	(0.062)	(0.060)	(0.061)	(0.061)
Peer CG1 went to uni	0.002	−0.007	−0.014	−0.014	−0.011
	(0.043)	(0.044)	(0.042)	(0.044)	(0.044)
Peer CG1 uni edu not known	0.008	0.008	0.002	0.006	−0.001
	(0.050)	(0.050)	(0.049)	(0.051)	(0.051)
Peer CG2 went to uni	0.030	−0.011	−0.019	−0.037	−0.022
	(0.043)	(0.042)	(0.041)	(0.042)	(0.041)
Peer CG2 uni edu not known	0.027	0.025	0.040	0.042	0.042
	(0.047)	(0.047)	(0.045)	(0.046)	(0.046)
Constant	−0.437***	−0.387***	−0.120	−0.052	−0.094
	(0.129)	(0.131)	(0.110)	(0.117)	(0.118)
Fem Peer: Fem vs Male Stu ($\beta_1 + \beta_3$)	0.488	0.467	0.195	0.221	0.213
Fem Peer: P-val	0.000	0.000	0.001	0.000	0.000
Classroom FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
R-squared	0.124	0.151	0.262	0.259	0.263
Observations	3189	3055	3009	2791	2762

Standard errors in parentheses and clustered at student level. * (p<0.10), ** (p<0.05), *** (p<0.01).

Table C.31: Peer-Assigned Score - Emotional Management

	(1)	(2)	(3)	(4)	(5)
Female Student (β_1)	0.027	0.022	−0.058	−0.021	−0.044
	(0.060)	(0.060)	(0.059)	(0.058)	(0.059)
Female Peer (β_2)	−0.223***	−0.156***	−0.158***	−0.190***	−0.231***
	(0.050)	(0.050)	(0.051)	(0.054)	(0.056)
Female Student x Female Peer (β_3)	0.270***	0.288***	0.310***	0.332***	0.357***
	(0.070)	(0.071)	(0.072)	(0.074)	(0.076)
Peer's Self Score		0.226***	0.215***	0.204***	0.192***
		(0.020)	(0.020)	(0.021)	(0.020)
Student's Teacher Score			0.277***	0.252***	0.281***
			(0.029)	(0.028)	(0.030)
Student's Self Score				0.101***	0.102***
				(0.029)	(0.030)
Peer's Teacher Score					0.081***
					(0.025)
Student born in Spain	0.284***	0.291***	0.248***	0.245**	0.240**
	(0.097)	(0.097)	(0.089)	(0.097)	(0.095)
Student Caregiver 1: Born in Spain	0.271***	0.275***	0.242***	0.275***	0.255***
	(0.085)	(0.086)	(0.080)	(0.090)	(0.092)
Student Caregiver 2: Born in Spain	−0.070	−0.057	−0.032	−0.081	−0.077
	(0.085)	(0.085)	(0.082)	(0.094)	(0.095)
Stu CG1 went to uni	−0.183***	−0.183***	−0.129**	−0.118**	−0.120**
	(0.060)	(0.061)	(0.057)	(0.059)	(0.060)
Stu CG1 uni edu not known	−0.226***	−0.233***	−0.163**	−0.183***	−0.183***
	(0.072)	(0.071)	(0.068)	(0.068)	(0.070)
Stu CG2 went to uni	0.221***	0.243***	0.170***	0.127**	0.119*
	(0.062)	(0.062)	(0.061)	(0.060)	(0.061)
Stu CG2 uni edu not known	−0.079	−0.077	−0.044	−0.070	−0.078
	(0.065)	(0.065)	(0.063)	(0.060)	(0.062)
Peer born in Spain	0.110*	0.095*	0.107*	0.093	0.076
	(0.056)	(0.056)	(0.058)	(0.062)	(0.063)
Peer Caregiver 1: Born in Spain	0.076	0.153***	0.158***	0.147**	0.113*
	(0.057)	(0.056)	(0.055)	(0.057)	(0.058)
Peer Caregiver 2: Born in Spain	−0.217***	−0.282***	−0.304***	−0.287***	−0.273***
	(0.063)	(0.062)	(0.062)	(0.066)	(0.067)
Peer CG1 went to uni	−0.011	−0.023	−0.019	−0.044	−0.048
	(0.043)	(0.044)	(0.045)	(0.047)	(0.049)
Peer CG1 uni edu not known	0.073	−0.016	−0.000	−0.005	−0.001
	(0.049)	(0.049)	(0.051)	(0.054)	(0.055)
Peer CG2 went to uni	0.050	−0.029	−0.029	−0.029	−0.032
	(0.045)	(0.045)	(0.046)	(0.048)	(0.050)
Peer CG2 uni edu not known	−0.086**	−0.049	−0.078*	−0.096**	−0.097**
	(0.043)	(0.043)	(0.045)	(0.047)	(0.047)
Constant	−0.277**	−0.266*	−0.232*	−0.153	−0.099
	(0.140)	(0.139)	(0.133)	(0.139)	(0.130)
Fem Peer: Fem vs Male Stu ($\beta_1 + \beta_3$)	0.297	0.310	0.252	0.311	0.312
Fem Peer: P-val	0.000	0.000	0.000	0.000	0.000
Classroom FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
R-squared	0.143	0.184	0.229	0.233	0.238
Observations	3003	2831	2592	2337	2239

Standard errors in parentheses and clustered at student level. * (p<0.10), ** (p<0.05), *** (p<0.01).

Table C.32: Peer-Assigned Score - Thinking Abilities

	(1)	(2)	(3)	(4)	(5)
Female Student (β_1)	−0.012	−0.024	-0.214^{***}	−0.109	−0.110
	(0.067)	(0.070)	(0.067)	(0.069)	(0.069)
Female Peer (β_2)	-0.337^{***}	-0.278^{***}	-0.273^{***}	-0.275^{***}	-0.205^{***}
	(0.059)	(0.062)	(0.062)	(0.065)	(0.067)
Female Student x Female Peer (β_3)	0.384^{***}	0.403^{***}	0.411^{***}	0.414^{***}	0.415^{***}
	(0.075)	(0.080)	(0.078)	(0.080)	(0.081)
Peer's Self Score		0.223^{***}	0.223^{***}	0.231^{***}	0.250^{***}
		(0.025)	(0.023)	(0.025)	(0.026)
Student's Teacher Score			0.429^{***}	0.382^{***}	0.370^{***}
			(0.037)	(0.036)	(0.036)
Student's Self Score				0.162^{***}	0.172^{***}
				(0.031)	(0.031)
Peer's Teacher Score					-0.145^{***}
					(0.030)
Student born in Spain	0.347^{***}	0.345^{***}	0.255^{**}	0.298^{***}	0.296^{***}
	(0.105)	(0.110)	(0.103)	(0.101)	(0.101)
Student Caregiver 1: Born in Spain	0.242^{***}	0.246^{***}	0.104	0.117	0.112
	(0.087)	(0.095)	(0.092)	(0.095)	(0.097)
Student Caregiver 2: Born in Spain	−0.037	−0.027	0.003	−0.041	−0.030
	(0.096)	(0.103)	(0.095)	(0.097)	(0.102)
Stu CG1 went to uni	-0.124^{*}	-0.122^{*}	-0.122^{*}	-0.124^{**}	-0.119^{*}
	(0.069)	(0.071)	(0.064)	(0.062)	(0.062)
Stu CG1 uni edu not known	-0.288^{***}	-0.282^{***}	-0.146^{*}	-0.132^{*}	-0.134^{*}
	(0.084)	(0.087)	(0.076)	(0.078)	(0.078)
Stu CG2 went to uni	0.175^{**}	0.159^{**}	0.066	0.066	0.080
	(0.072)	(0.073)	(0.062)	(0.064)	(0.065)
Stu CG2 uni edu not known	−0.007	−0.021	−0.053	−0.029	−0.018
	(0.080)	(0.082)	(0.068)	(0.069)	(0.069)
Peer born in Spain	0.013	−0.012	−0.050	−0.026	0.001
	(0.054)	(0.054)	(0.052)	(0.059)	(0.059)
Peer Caregiver 1: Born in Spain	0.070	0.057	0.071	0.082	0.127^{*}
	(0.063)	(0.067)	(0.063)	(0.066)	(0.066)
Peer Caregiver 2: Born in Spain	-0.137^{*}	−0.085	-0.131^{*}	−0.128	-0.142^{*}
	(0.072)	(0.077)	(0.076)	(0.078)	(0.079)
Peer CG1 went to uni	0.115^{***}	0.080^{*}	0.103^{**}	0.110^{**}	0.096^{**}
	(0.044)	(0.046)	(0.044)	(0.046)	(0.047)
Peer CG1 uni edu not known	0.218^{***}	0.124^{**}	0.158^{***}	0.150^{***}	0.086
	(0.051)	(0.052)	(0.049)	(0.053)	(0.056)
Peer CG2 went to uni	−0.028	−0.044	-0.093^{**}	-0.069^{*}	−0.031
	(0.044)	(0.044)	(0.040)	(0.042)	(0.042)
Peer CG2 uni edu not known	-0.174^{***}	-0.114^{**}	-0.161^{***}	-0.138^{***}	-0.119^{**}
	(0.048)	(0.049)	(0.045)	(0.048)	(0.048)
Constant	-0.291^{**}	-0.306^{**}	−0.040	−0.155	−0.222
	(0.147)	(0.148)	(0.148)	(0.157)	(0.157)
Fem Peer: Fem vs Male Stu ($\beta_1 + \beta_3$)	0.371	0.379	0.197	0.306	0.305
Fem Peer: P-val	0.000	0.000	0.008	0.000	0.000
Classroom FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
R-squared	0.172	0.196	0.291	0.297	0.303
Observations	2737	2479	2359	2097	2065

Standard errors in parentheses and clustered at student level. * (p<0.10), ** (p<0.05), *** (p<0.01).

Table C.33: Peer-Assigned Scores with Female Tch: All 5 dimensions

	Autonomy	Cooperation	Responsibility	Emotion	Thinking
Female Student (β_1)	-0.075	-0.115**	-0.023	-0.049	-0.137**
	(0.069)	(0.058)	(0.060)	(0.060)	(0.066)
Female Peer (β_2)	-0.226***	-0.345***	-0.191***	-0.230***	-0.254***
	(0.066)	(0.055)	(0.059)	(0.056)	(0.069)
Female Student x Female Peer (β_3)	0.454***	0.461***	0.287***	0.336***	0.470***
	(0.091)	(0.074)	(0.075)	(0.077)	(0.083)
Peer's Self Score	0.182***	0.270***	0.199***	0.196***	0.238***
	(0.027)	(0.021)	(0.022)	(0.021)	(0.026)
Teacher Score	0.570***	0.367***	0.383***	0.316***	0.395***
	(0.079)	(0.080)	(0.062)	(0.082)	(0.107)
Teacher Female × Teacher Score	-0.272***	-0.080	0.005	-0.056	-0.025
	(0.084)	(0.085)	(0.071)	(0.086)	(0.113)
Student's Self Score	0.132***	0.089***	0.095***	0.107***	0.170***
	(0.040)	(0.027)	(0.030)	(0.030)	(0.031)
Peer's Teacher Score	-0.105***	-0.014	-0.067**	0.063**	-0.142***
	(0.029)	(0.027)	(0.026)	(0.025)	(0.032)
$\beta_1 + \beta_3$	0.379***	0.346***	0.265***	0.287***	0.333***
Observations	1721	2389	2292	2112	1707
Adj. R-squared	0.258	0.245	0.243	0.205	0.302

Note: All regressions control for student and peer background characteristics, and school fixed effects. Standard errors clustered at student level in parentheses. The dependent variable is peer-assigned score. *p<0.10, ** p<0.05, *** p<0.01.

Table C.34: Autonomy: Separate Regressions for Peer FE and Student FE

	Peer FEs	**Peer FEs**	**Peer FEs**	**Student FEs**	**Student FEs**	**Student FEs**
Fem Student × Fem Peer	0.320***	0.323***		0.272***	0.399***	
	(0.074)	(0.091)		(0.063)	(0.101)	
Male Student × Male Peer	0.003		0.323***	0.127		0.399***
	(0.076)		(0.091)	(0.081)		(0.101)
Female Student		-0.003				
		(0.076)				
Male Student			-0.320***			
			(0.074)			
Female Peer					-0.127	
					(0.081)	
Male Peer						-0.272***
						(0.063)
Peer Controls	No	No	No	Yes	Yes	Yes
Student Controls	Yes	Yes	Yes	No	No	No
Observations	2264	2264	2264	2373	2373	2373
Adj. R-squared	0.524	0.524	0.524	0.574	0.574	0.574

Note: Columns 1–3 (Peer FE) report regressions of the Autonomy score on student×peer gender interaction and main effects, controlling for the Student's Teacher score, Self score, and background characteristics; standard errors are clustered at the student level. Columns 4–6 (Student FE) report analogous regressions with all controls at the peer level, also clustering SEs by student. *p<0.10, ** p<0.05, *** p<0.010. Standard errors in parentheses.

Table C.35: Robustness Check with EO Scores - Peer-Assigned Score: Peer FE vs. Student FE Across 5 Domains

	Autonomy		**Cooperation**		**Responsibility**		**Emotion Mngt.**		**Thinking Ab.**	
	Peer FE	**Student FE**	**Peer FE**	**Student FE**	**Peer FE**	**Student FE**	**Peer FE**	**Student FE**	**Peer FE**	**Student FE**
Fem Student x Fem Peer	0.363***	0.136	0.341***	0.078	0.328***	0.013	0.276**	0.173	0.464***	0.084
	(0.104)	(0.119)	(0.088)	(0.101)	(0.103)	(0.095)	(0.114)	(0.110)	(0.144)	(0.152)
Male Student x Male Peer	0.030	0.091	-0.048	0.243**	-0.227**	0.086	-0.152	-0.164	-0.010	0.333**
	(0.135)	(0.131)	(0.104)	(0.106)	(0.101)	(0.101)	(0.110)	(0.120)	(0.150)	(0.160)
EO Score	0.183**	-0.062	0.152**	0.022	0.226***	-0.028	0.083	-0.053	0.131	-0.012
	(0.075)	(0.040)	(0.069)	(0.048)	(0.054)	(0.047)	(0.058)	(0.054)	(0.080)	(0.067)
Self Score	0.131*	0.153***	0.112**	0.245***	0.118**	0.222***	0.204***	0.283***	0.310***	0.228***
	(0.076)	(0.046)	(0.054)	(0.042)	(0.056)	(0.050)	(0.067)	(0.045)	(0.066)	(0.067)
Peer Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Student Controls	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
Observations	1068	1127	1435	1519	1590	1691	1153	1253	892	964
Adj. R-squared	0.644	0.453	0.601	0.359	0.598	0.400	0.594	0.301	0.646	0.367

Note: Standard errors in parentheses and clustered at the student level. The dependent variable is peer-assigned score. Under Peer FEs, the EO and Self Score refers to student's scores and under Student FEs, these refer to the peer's scores . *p<0.10, ** p<0.05, *** p<0.010.

Table C.36: Lab-in-the-field: Peer FE vs. Student FE with Peer-Assigned Score

	Autonomy		**Cooperation**		**Responsibility**		**Emotion Mngt.**		**Thinking Ab.**	
	Peer FE	**Student FE**	**Peer FE**	**Student FE**	**Peer FE**	**Student FE**	**Peer FE**	**Student FE**	**Peer FE**	**Student FE**
Fem Student x Fem Peer	0.184***	0.171**	0.327***	0.114	0.291***	0.058	0.054	0.081	0.211***	0.124
	(0.071)	(0.077)	(0.076)	(0.079)	(0.069)	(0.077)	(0.071)	(0.087)	(0.078)	(0.085)
Male Student x Male Peer	0.047	-0.044	-0.049	0.145*	-0.016	0.144*	-0.085	0.037	0.032	0.080
	(0.064)	(0.078)	(0.061)	(0.080)	(0.064)	(0.079)	(0.064)	(0.086)	(0.063)	(0.080)
EO Score	0.304***	-0.055	0.278***	0.043	0.319***	-0.028	0.195***	0.036	0.315***	-0.041
	(0.030)	(0.034)	(0.043)	(0.043)	(0.040)	(0.043)	(0.039)	(0.048)	(0.038)	(0.045)
Self Score	0.135***	0.354***	0.074**	0.453***	0.098**	0.439***	0.044	0.313***	0.135***	0.377***
	(0.038)	(0.042)	(0.036)	(0.045)	(0.040)	(0.047)	(0.034)	(0.040)	(0.034)	(0.038)
Peer Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Student Controls	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
Observations	2093	2112	2090	2109	2095	2114	2073	2093	2076	2098
R-squared	0.671	0.689	0.703	0.687	0.719	0.690	0.676	0.626	0.699	0.662

Note: Standard errors in parentheses and clustered at the student level. The dependent variable is peer-assigned score. Under Peer FEs, the EO and Self Score refers to student's scores and under Student FEs, these refer to the peer's scores . *p<0.10, ** p<0.05, *** p<0.010.

Robustness: Descriptive statistics of the lab-in-the-field survey

Table C.37: Summary Statistics: Lab-in-the-field Self-Evaluation

	Female			**Male**			**Total**				
	Mean	**SD**	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**	**Min**	**Max**	**N**
Autonomy	3.9	.78	494	3.84	.84	515	3.86	.81	1	5	1044
Cooperation	3.99	.78	496	3.86	.84	514	3.92	.81	1	5	1045
Responsibility	4.25	.74	498	4.12	.76	514	4.17	.77	1	5	1047
Emt. Mngt.	3.72	.85	498	3.8	.86	512	3.76	.85	1	5	1045
Thk. Abi.	3.65	.9	494	3.74	.89	514	3.69	.9	1	5	1043

Table C.38: Unequal variance t-test (Female vs. Male): Lab-in-the-field Self-Evaluation

Variable	Diff	SE	p-value
Autonomy	0.06	0.05	0.218
Cooperation	0.13**	0.05	0.011
Responsibility	0.12***	0.05	0.008
Emt. Mngt.	-0.09	0.05	0.109
Thk. Abi.	-0.09*	0.06	0.095

Table C.39: Summary Statistics: Lab-in-the-field Peer Evaluations

	Female Students			**Male Students**			**Total**				
	Mean	**SD**	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**	**Min**	**Max**	**N**
Peer Evalrs/ Student	2.79	0.53	488	2.78	0.55	510	2.78	0.54	1.00	5.00	1037
Panel A: All Peers											
Autonomy	4.01	0.87	1342	3.84	0.98	1408	3.91	0.94	1.00	5.00	2858
Cooperation	3.98	0.89	1340	3.77	0.99	1407	3.86	0.95	1.00	5.00	2853
Responsibility	4.18	0.81	1342	3.89	1.01	1408	4.02	0.94	1.00	5.00	2858
Emt. Mngt.	3.89	0.88	1331	3.77	0.97	1398	3.82	0.93	1.00	5.00	2834
Thk. Abi.	3.89	0.91	1336	3.75	1.00	1392	3.81	0.97	1.00	5.00	2833
Panel B: Female Peers											
Autonomy	4.10	0.83	674	3.83	0.98	633	3.95	0.93	1.00	5.00	1369
Cooperation	4.07	0.86	674	3.73	1.00	632	3.89	0.95	1.00	5.00	1366
Responsibility	4.23	0.81	673	3.88	1.02	632	4.05	0.94	1.00	5.00	1367
Emt. Mngt.	3.93	0.85	667	3.77	0.91	626	3.85	0.89	1.00	5.00	1352
Thk. Abi.	3.95	0.89	671	3.70	1.02	623	3.81	0.97	1.00	5.00	1354
Panel C: Male Peers											
Autonomy	3.90	0.92	615	3.84	0.99	735	3.86	0.96	1.00	5.00	1392
Cooperation	3.89	0.91	614	3.78	0.99	735	3.82	0.96	1.00	5.00	1391
Responsibility	4.11	0.83	616	3.88	1.02	736	3.98	0.94	1.00	5.00	1394
Emt. Mngt.	3.84	0.91	614	3.74	1.02	733	3.78	0.97	1.00	5.00	1389
Thk. Abi.	3.83	0.93	613	3.79	0.99	730	3.80	0.97	1.00	5.00	1385

Table C.40: Unequal variance t-tests (Female vs. Male Students): Lab-in-the-field Peer Evaluations

	Diff	SE	p-value
Panel A: All Peers			
Autonomy	0.17***	0.04	0.000
Cooperation	0.21***	0.04	0.000
Responsibility	0.29***	0.03	0.000
Emt. Mngt.	0.12***	0.04	0.000
Thk. Abi.	0.14***	0.04	0.000
Panel B: Female Peers			
Autonomy	0.27***	0.05	0.000
Cooperation	0.34***	0.05	0.000
Responsibility	0.36***	0.05	0.000
Emt. Mngt.	0.16***	0.05	0.001
Thk. Abi.	0.25***	0.05	0.000
Panel C: Male Peers			
Autonomy	0.07	0.05	0.205
Cooperation	0.10**	0.05	0.044
Responsibility	0.23***	0.05	0.000
Emt. Mngt.	0.09*	0.05	0.075
Thk. Abi.	0.04	0.05	0.482

Table C.41: Summary Statistics: Lab-in-the-field EO Evaluations

	Female Students			Male Students			Total				
	Mean	**SD**	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**	**Min**	**Max**	**N**
EOs/Student	1.06	0.24	440	1.06	0.24	447	1.07	0.25	1.00	2.00	919
Panel A: All EO											
Autonomy	3.12	0.92	466	2.93	0.96	474	3.02	0.94	1.00	5.00	975
Cooperation	2.89	0.74	467	2.74	0.83	474	2.82	0.79	1.00	5.00	976
Responsibility	3.77	0.72	467	3.41	1.01	474	3.58	0.90	1.00	5.00	978
Emotional Management	3.13	0.77	465	3.00	0.86	468	3.07	0.81	1.00	5.00	968
Thinking Abilities	2.84	0.78	466	2.67	0.80	474	2.75	0.79	1.00	5.00	975
Panel B: Female EO											
Autonomy	3.13	0.96	352	2.98	0.96	362	3.04	0.97	1.00	5.00	739
Cooperation	2.85	0.78	353	2.80	0.86	362	2.83	0.82	1.00	5.00	740
Responsibility	3.81	0.73	353	3.47	1.01	362	3.63	0.90	1.00	5.00	742
Emotional Management	3.19	0.76	351	3.02	0.86	356	3.11	0.82	1.00	5.00	732
Thinking Abilities	2.81	0.80	352	2.72	0.80	362	2.76	0.80	1.00	4.50	739
Panel C: Male EO											
Autonomy	3.12	0.80	114	2.75	0.91	112	2.94	0.86	1.00	5.00	236
Cooperation	3.01	0.59	114	2.55	0.71	112	2.79	0.69	1.00	4.50	236
Responsibility	3.64	0.68	114	3.19	0.98	112	3.41	0.87	1.00	5.00	236
Emotional Management	2.96	0.77	114	2.92	0.84	112	2.95	0.80	1.00	5.00	236
Thinking Abilities	2.92	0.72	114	2.50	0.75	112	2.71	0.75	1.00	5.00	236

Table C.42: Welch Unequal-Variance t-tests (Female vs. Male) for Lab-in-the-field EO Evaluations

	Diff (F–M)	SE	p-value
Panel A: All EO			
Autonomy	0.20***	0.06	0.001
Cooperation	0.15***	0.05	0.003
Responsibility	0.36***	0.06	0.000
Emotional Management	0.13**	0.05	0.012
Thinking Abilities	0.17***	0.05	0.001
Panel B: Female EO			
Autonomy	0.15**	0.07	0.042
Cooperation	0.05	0.06	0.374
Responsibility	0.34***	0.07	0.000
Emotional Management	0.16***	0.06	0.007
Thinking Abilities	0.09	0.06	0.145
Panel C: Male EO			
Autonomy	0.36***	0.11	0.002
Cooperation	0.46***	0.09	0.000
Responsibility	0.45***	0.11	0.000
Emotional Management	0.04	0.11	0.703
Thinking Abilities	0.42***	0.10	0.000

Robustness: BESSI Descriptive Statistics

Table C.43: BESSI Summary Statistics

	Self			**Peer**			**Teacher**		
	Mean	**SD**	**N**	**Mean**	**SD**	**N**	**Mean**	**SD**	**N**
Self Management	3.48	0.68	643	3.35	0.85	1787	3.21	1.00	553
Social Engagement	3.20	0.78	643	3.31	0.83	1787	3.13	0.87	552
Cooperation	3.65	0.65	643	3.43	0.81	1786	3.36	0.74	552
Emotional Management	3.21	0.81	643	3.33	0.80	1784	3.31	0.77	552
Innovation	3.35	0.72	643	3.30	0.82	1781	3.25	0.77	553

Robustness: BESSI Self-assigned Score

Table C.44: BESSI Self-Assigned Score Treated Group (Endline) - Student's Self Score

	(1)	(2)	(3)
Female Student	-0.213^{***}	-0.300^{***}	-0.309^{***}
	(0.066)	(0.069)	(0.069)
Student's Teacher Score		0.348^{***}	0.298^{***}
		(0.044)	(0.051)
Student's Avg Peer Score			0.148^{**}
			(0.072)
Student Background Controls	*Yes*	*Yes*	*Yes*
Classroom FE	*Yes*	*Yes*	*Yes*
R-squared (adj.)	0.034	0.137	0.146
Observations	596	509	509

Note: Robust standard errors are in parenthesis. *p<0.10, **p<0.05, ***p<0.01.

Table C.45: BESSI Self-Assigned Score Treated Group (Endline) - Student's Self Score

	(1)	(2)	(3)
Female Student	−0.085	−0.146**	−0.162***
	(0.055)	(0.061)	(0.060)
Student's Teacher Score		0.213***	0.169***
		(0.047)	(0.053)
Student's Avg Peer Score			0.132**
			(0.063)
Student Background Controls	*Yes*	*Yes*	*Yes*
Classroom FE	*Yes*	*Yes*	*Yes*
R-squared (adj.)	0.012	0.055	0.065
Observations	596	509	509

Note: Robust standard errors are in parenthesis. *p<0.10, **p<0.05, ***p<0.01.

Table C.46: BESSI Self-Assigned Score Treated Group (Endline) - Student's Self Score

	(1)	(2)	(3)
Female Student	−0.420***	−0.459***	−0.467***
	(0.067)	(0.072)	(0.073)
Student's Teacher Score		0.207***	0.180***
		(0.052)	(0.062)
Student's Avg Peer Score			0.071
			(0.081)
Student Background Controls	*Yes*	*Yes*	*Yes*
Classroom FE	*Yes*	*Yes*	*Yes*
R-squared (adj.)	0.065	0.097	0.097
Observations	596	509	509

Note: Robust standard errors are in parenthesis. *p<0.10, **p<0.05, ***p<0.01.

Table C.47: BESSI Self-Assigned Score Treated Group (Endline) - Student's Self Score

	(1)	(2)	(3)
Female Student	-0.075	-0.220^{***}	-0.229^{***}
	(0.059)	(0.063)	(0.064)
Student's Teacher Score		0.232^{***}	0.199^{***}
		(0.035)	(0.047)
Student's Avg Peer Score			0.071
			(0.073)
Student Background Controls	*Yes*	*Yes*	*Yes*
Classroom FE	*Yes*	*Yes*	*Yes*
R-squared (adj.)	-0.008	0.078	0.079
Observations	596	510	510

Note: Robust standard errors are in parenthesis. *p<0.10, **p<0.05, ***p<0.01.

Table C.48: BESSI Self-Assigned Score Treated Group (Endline) - Student's Self Score

	(1)	(2)	(3)
Female Student	0.135^{**}	0.086	0.060
	(0.063)	(0.068)	(0.069)
Student's Teacher Score		0.222^{***}	0.155^{***}
		(0.048)	(0.060)
Student's Avg Peer Score			0.139^{**}
			(0.067)
Student Background Controls	*Yes*	*Yes*	*Yes*
Classroom FE	*Yes*	*Yes*	*Yes*
R-squared (adj.)	-0.001	0.048	0.056
Observations	596	510	510

Note: Robust standard errors are in parenthesis. *p<0.10, **p<0.05, ***p<0.01.

Robustness: BESSI Peer-assigned Score

Table C.49: BESSI Peer-Assigned Score (Endline): Full specification

	Self Mngt	Social Engt	Cooperation	Emotional Mngt	Innovation
Female Student	-0.047	-0.078	-0.024	-0.055	0.022
	(0.078)	(0.078)	(0.068)	(0.081)	(0.075)
Female Peer	-0.296***	-0.130	-0.189**	-0.251***	-0.095
	(0.074)	(0.081)	(0.073)	(0.081)	(0.076)
Female Student × Female Peer	0.413***	0.281**	0.337***	0.385***	0.262**
	(0.112)	(0.120)	(0.107)	(0.115)	(0.111)
Peer's Self Score	0.288***	0.227***	0.437***	0.248***	0.243***
	(0.045)	(0.042)	(0.051)	(0.043)	(0.041)
Student's Teacher Score	0.442***	0.317***	0.326***	0.367***	0.477***
	(0.035)	(0.043)	(0.041)	(0.044)	(0.043)
Student's Self Score	0.038	0.095**	0.165***	0.079*	0.104***
	(0.048)	(0.044)	(0.049)	(0.041)	(0.036)
Peer's Teacher Score	-0.090***	-0.050	-0.015	-0.054	-0.076*
	(0.031)	(0.040)	(0.047)	(0.044)	(0.039)
$\beta_1 + \beta_3$	0.366***	0.203**	0.312***	0.330***	0.284***
Male × Male Mean	3.36	3.35	3.46	3.38	3.19
Observations	743	739	738	738	741
Adj. R-squared	0.402	0.211	0.292	0.252	0.307

Note: All regressions control for student and peer background characteristics, and classroom fixed effects. Standard errors are in parentheses and clustered at student level. The dependent variable is peer-assigned score. *p<0.10, ** p<0.05, *** p<0.01.

Table C.50: BESSI Peer-Assigned Score (Endline) - Social Engagement

	(1)	(2)	(3)	(4)	(5)
Female Student	−0.009	−0.025	−0.045	−0.021	−0.078
	(0.056)	(0.055)	(0.057)	(0.057)	(0.080)
Female Peer	−0.267***	−0.222***	−0.192***	−0.202***	−0.130
	(0.058)	(0.057)	(0.057)	(0.057)	(0.082)
Female Student × Female Peer	0.303***	0.316***	0.263***	0.278***	0.281**
	(0.083)	(0.082)	(0.083)	(0.083)	(0.123)
Peer's Self Score		0.241***	0.231***	0.231***	0.227***
		(0.029)	(0.029)	(0.029)	(0.042)
Student's Teacher Score			0.336***	0.303***	0.317***
			(0.028)	(0.030)	(0.042)
Student's Self Score				0.101***	0.095**
				(0.031)	(0.042)
Peer's Teacher Score					−0.050
					(0.041)
Constant	3.126***	2.362***	1.570***	1.315***	1.628***
	(0.109)	(0.137)	(0.159)	(0.181)	(0.275)
Student Background Controls	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Peer Background Controls	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Classroom FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
R-squared	0.128	0.172	0.277	0.284	0.303
Observations	1544	1544	1292	1290	739

Note: Standard errors are in parentheses and clustered at student level. The dependent variable is peer-assigned score. *p<0.10, **p<0.05, ***p<0.01.

Table C.51: BESSI Peer-Assigned Score (Endline) - Cooperation

	(1)	(2)	(3)	(4)	(5)
Female Student	0.062	0.035	0.015	0.026	−0.024
	(0.056)	(0.053)	(0.056)	(0.055)	(0.072)
Female Peer	-0.266^{***}	-0.228^{***}	-0.186^{***}	-0.194^{***}	-0.189^{**}
	(0.058)	(0.054)	(0.055)	(0.055)	(0.075)
Female Student × Female Peer	0.332^{***}	0.343^{***}	0.255^{***}	0.263^{***}	0.337^{***}
	(0.081)	(0.077)	(0.080)	(0.080)	(0.110)
Peer's Self Score		0.395^{***}	0.380^{***}	0.392^{***}	0.437^{***}
		(0.034)	(0.034)	(0.034)	(0.049)
Student's Teacher Score			0.361^{***}	0.337^{***}	0.326^{***}
			(0.031)	(0.032)	(0.044)
Student's Self Score				0.132^{***}	0.165^{***}
				(0.034)	(0.048)
Peer's Teacher Score					−0.015
					(0.046)
Constant	3.317^{***}	1.905^{***}	0.882^{***}	0.429^{**}	0.551
	(0.110)	(0.154)	(0.179)	(0.218)	(0.363)
Student Background Controls	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Peer Background Controls	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Classroom FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
R-squared	0.138	0.226	0.314	0.323	0.375
Observations	1543	1543	1291	1289	738

Note: Standard errors are in parentheses and clustered at student level. The dependent variable is peer-assigned score. *p<0.10, **p<0.05, ***p<0.01.

Table C.52: BESSI Peer-Assigned Score (Endline) - Emotional Management

	(1)	(2)	(3)	(4)	(5)
Female Student	0.028	0.024	−0.013	0.030	−0.055
	(0.058)	(0.056)	(0.059)	(0.060)	(0.083)
Female Peer	-0.343^{***}	-0.235^{***}	-0.238^{***}	-0.241^{***}	-0.251^{***}
	(0.059)	(0.058)	(0.058)	(0.058)	(0.081)
Female Student × Female Peer	0.318^{***}	0.319^{***}	0.324^{***}	0.334^{***}	0.385^{***}
	(0.082)	(0.079)	(0.081)	(0.080)	(0.116)
Peer's Self Score		0.250^{***}	0.237^{***}	0.241^{***}	0.248^{***}
		(0.029)	(0.029)	(0.029)	(0.041)
Student's Teacher Score			0.366^{***}	0.345^{***}	0.367^{***}
			(0.030)	(0.031)	(0.044)
Student's Self Score				0.101^{***}	0.079^{**}
				(0.028)	(0.039)
Peer's Teacher Score					−0.054
					(0.044)
Constant	3.466^{***}	2.620^{***}	1.518^{***}	1.220^{***}	1.680^{***}
	(0.110)	(0.144)	(0.171)	(0.186)	(0.316)
Student Background Controls	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Peer Background Controls	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Classroom FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
R-squared	0.123	0.174	0.291	0.298	0.339
Observations	1542	1542	1290	1288	738

Note: Standard errors are in parentheses and clustered at student level. The dependent variable is peer-assigned score. *p<0.10, **p<0.05, ***p<0.01.

Table C.53: BESSI Peer-Assigned Score (Endline) - Self Management

	(1)	(2)	(3)	(4)	(5)
Female Student	0.237***	0.229***	−0.012	−0.005	−0.047
	(0.058)	(0.057)	(0.057)	(0.058)	(0.076)
Female Peer	−0.385***	−0.380***	−0.364***	−0.366***	−0.296***
	(0.059)	(0.058)	(0.055)	(0.055)	(0.078)
Female Student × Female Peer	0.403***	0.401***	0.357***	0.363***	0.413***
	(0.083)	(0.082)	(0.078)	(0.078)	(0.112)
Peer's Self Score		0.217***	0.229***	0.230***	0.288***
		(0.034)	(0.032)	(0.032)	(0.044)
Student's Teacher Score			0.451***	0.438***	0.442***
			(0.022)	(0.023)	(0.033)
Student's Self Score				0.062*	0.038
				(0.033)	(0.044)
Peer's Teacher Score					−0.090***
					(0.031)
Constant	3.182***	2.452***	1.279***	1.095***	1.379***
	(0.113)	(0.161)	(0.158)	(0.184)	(0.273)
Student Background Controls	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Peer Background Controls	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Classroom FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
R-squared	0.184	0.211	0.429	0.431	0.472
Observations	1544	1544	1294	1292	743

Note: Standard errors are in parentheses and clustered at student level. The dependent variable is peer-assigned score. *p<0.10, **p<0.05, ***p<0.01.

Table C.54: BESSI Peer-Assigned Score (Endline) - Innovation

	(1)	(2)	(3)	(4)	(5)
Female Student	0.271***	0.245***	0.146**	0.120**	0.022
	(0.058)	(0.056)	(0.057)	(0.058)	(0.076)
Female Peer	−0.133**	−0.176***	−0.140**	−0.162***	−0.095
	(0.060)	(0.058)	(0.058)	(0.057)	(0.077)
Female Student × Female Peer	0.212**	0.231***	0.212***	0.240***	0.262**
	(0.084)	(0.082)	(0.080)	(0.080)	(0.113)
Peer's Self Score		0.267***	0.231***	0.241***	0.243***
		(0.030)	(0.030)	(0.030)	(0.042)
Student's Teacher Score			0.467***	0.443***	0.477***
			(0.029)	(0.030)	(0.041)
Student's Self Score				0.122***	0.104***
				(0.028)	(0.035)
Peer's Teacher Score					−0.076*
					(0.040)
Constant	3.124***	2.258***	1.039***	0.669***	1.045***
	(0.111)	(0.148)	(0.162)	(0.180)	(0.267)
Student Background Controls	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Peer Background Controls	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Classroom FE	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
R-squared	0.146	0.194	0.346	0.356	0.387
Observations	1539	1539	1290	1288	741

Note: Standard errors are in parentheses and clustered at student level. The dependent variable is peer-assigned score. *p<0.10, **p<0.05, ***p<0.01.

Appendix D

Photos of the Standardized Activity

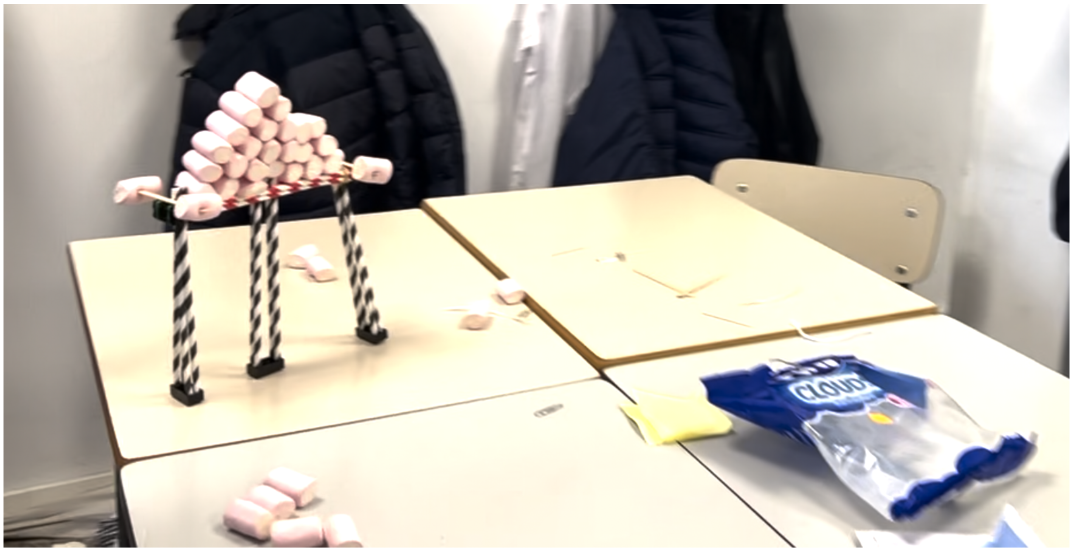


Figure D.1: Students constructing marshmallow towers

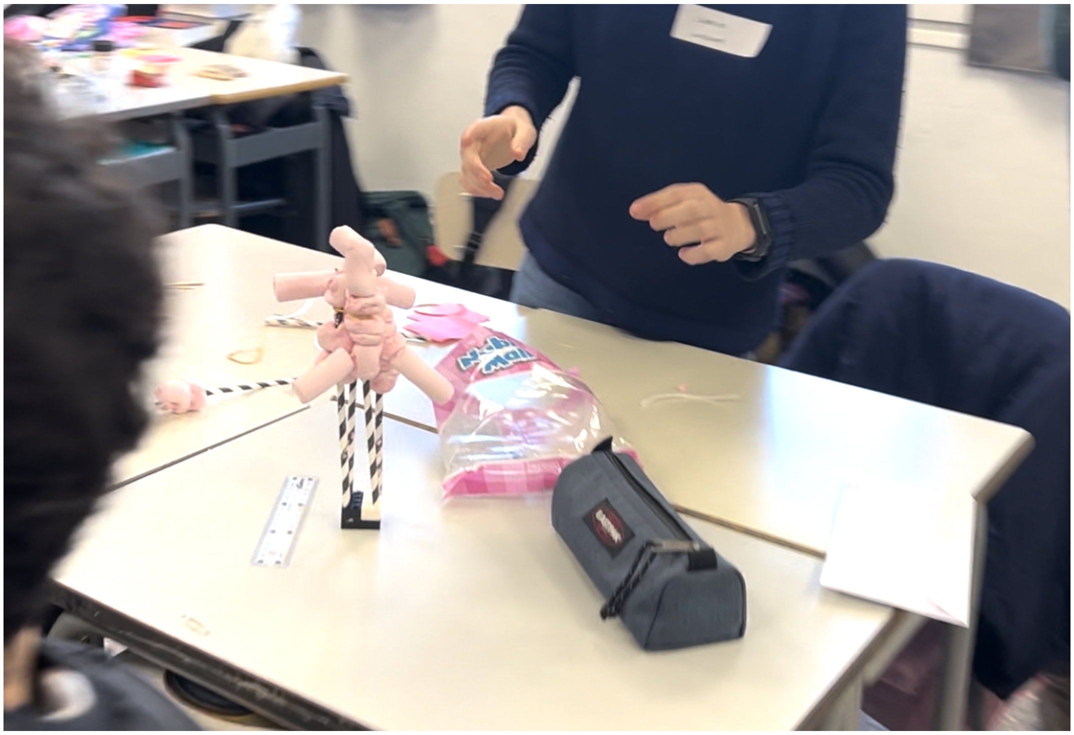


Figure D.2: Students constructing marshmallow towers

GEAR

GRADUATE
PROGRAM
IN APPLIED
ECONOMIC
RESEARCH

UAB Universitat Autònoma
de Barcelona