Journal of Development Economics

Developing Socioemotional Skills in Early Adolescence: Experimental Evidence from Morocco's Public Schools

--Manuscript Draft--

Manuscript Number:	DEVEC-D-25-00596R1
Article Type:	Registered Report Stage 1: Proposal
Section/Category:	Health, Education, gender, poverty
Keywords:	Education; Human capital; Morocco; socioemotional
Corresponding Author:	Andreas de Barros, PhD University of California Irvine UNITED STATES
First Author:	Andreas de Barros
Order of Authors:	Andreas de Barros
	Alejandra Campos Quintero
	Paul Glewwe
	Nikhil Kumar
	Laure Lépine
Abstract:	We experimentally evaluate the effectiveness of a socioemotional support intervention in public schools in Morocco. Implemented as part of a broader school reform, the intervention trains and equips social specialists in schools to give workshops in small classes for students in the first and second grades of lower secondary education. The intervention aims to develop students' intrapersonal and interpersonal skills, and ultimately improve student learning and reduce student dropout. Our study contributes to the evidence base on school-based socioemotional interventions in low- and lower-middle-income countries.
Response to Reviewers:	Thank you very much for your helpful comments. We're submitting our responses in the attached letter.

Pre-Results Review: Stage 1 Submission

Developing Socioemotional Skills in Early Adolescence: Experimental Evidence from Morocco's Public Schools*

Author details removed from title page

13 June 2025

Abstract

We experimentally evaluate the effectiveness of a socioemotional support intervention in public schools in Morocco. Implemented as part of a broader school reform, the intervention trains and equips social specialists in schools to give workshops in small classes for students in the first and second grades of lower secondary education. The intervention aims to develop students' intrapersonal and interpersonal skills, and ultimately improve student learning and reduce student dropout. Our study contributes to the evidence base on school-based socioemotional interventions in low- and lower-middle-income countries.

Keywords: Education; human capital; Morocco; socioemotional.

JEL classification: I20, I21, I28, O15, O22.

*This document follows the "reporting checklist" of the Journal of Development Economics (JDE) pre-results review process (Stage 1). We thank the Ministry of National Education, Preschools and Sports in Morocco for supporting both this study and a separate impact evaluation of the Pioneer School Program in public lower-secondary schools. We gratefully acknowledge generous funding from the Ministry. We are grateful to the Morocco Innovation and Evaluation Lab (MEL) and J-PAL staff—Youssef Assarssah, Najiba El Amrani Mida, Florencia Devoto, Sara Hassani, Andrea Salem, and Zakaria Mansouri—for their excellent field coordination, research assistance, and project leadership. We thank Mathilde Col and Quentin Daviot of EVAL-LAB for their excellent support in the data collection process. Data protection was secured in compliance with the relevant local legal provisions and regulations concerning Human Subjects research at X University, Y University, and Z University. The order of authors is alphabetical; co-authorship is shared equally among the authors.

Timeline: The study's baseline was conducted in September 2024. The study's endline will be conducted from 10 June 2025 to 13 June 2025; after data entry and cleaning, the co-authors will have access to endline data from 26 June 2025. Our expected date for completion of the pre-specified research design and disposition of the Stage 2 report is 15 January 2026.

1 Introduction

1.1 Background and relevance of the study

The importance of socioemotional skills as educational outcomes is increasingly recognized. While socioemotional outcomes are often seen as notable predictors of students' academic success, many education systems have begun to embrace their importance as a goal in itself. This shift is informed by research that connects socioemotional skills to later life outcomes, including labor market outcomes (Deming, 2017). It also rests on research showing that teachers' ability to promote socioemotional skills is largely orthogonal to their test-score value added Jackson (2018). However, the technology for producing these skills is not well understood—little is known about how public school systems can promote them with targeted interventions, and most of the related research focuses on upper-middle-income and high-income contexts.

Our study addresses this gap by studying the Moroccan government's efforts to embed a targeted socioemotional support intervention within a larger reform of the country's public lower secondary schools. The intervention trains and equips social specialists working in these schools to give four workshops per year in small classes for students in the first and second years of lower secondary education (grades 7 and 8). Beyond these workshops, the intervention intends to promote awareness of the social specialists working in the schools and build relationships of trust between the specialists and the students. The intervention thus aims to develop students' intrapersonal and interpersonal skills, and ultimately improve student learning and reduce student dropout.

The study involves 200 public lower secondary schools, all of which participate in the first year of the school reform. Through a cluster-randomized trial across these 200 schools, we estimate the causal effect of the socioemotional support intervention. We focus on the intervention's effects on socioemotional skills, yet we also investigate broader effects on student learning and dropping out. In secondary analyses, we explore whether subgroups of students are affected differentially. In addition, we use data on intermediate outcomes and mechanisms to pinpoint where, in the case of a null finding, the intervention's Theory of Change may not have operated as expected.

The study's first and most direct contribution is to the nascent literature on government-led, school-based interventions that aim to promote socioemotional skills in low- and lower-income countries.[1] Prior work in this area has studied school-based interventions

[1]For a review and meta-analysis for high-income countries, see Wilson et al. (2025) and Cipriano et al. (2023). For related work from upper-middle-income countries, see Alan et al. (2019), Ganimian (2020), and Santos et al. (2022). For a study on the effects of school guidance counselors in the United States, see Mulhern (2023).

delivered by researchers and NGOs in Central America (Dinarte-Diaz et al., 2024), India (Edmonds et al., 2023; Dhar et al., 2022), and Zambia (Ashraf et al., 2020).[2] To our best knowledge, this is the first large-scale, experimental evaluation of a government-led, socioemotional skill intervention implemented in public schools in a low- or lower-middle-income country.[3]

By focusing on early adolescence (grades 7 and 8), our study examines a critical period for youth development, during which students experience significant physical, cognitive, emotional, and social transitions. This stage is characterized by increasing autonomy, the reconfiguration of peer and adult relationships, and heightened sensitivity to social dynamics—which shape the development of key socioemotional competencies. While much of the (limited) existing research on the production of socioemotional skills in low- and lower-middle-income countries has focused on early-childhood development (e.g., Dillon et al., 2017), primary-grade interventions (e.g., Barrera-Osorio et al., 2024), or late adolescence (e.g., Beaman et al., 2021; Krishnan and Krutikova, 2013), our research provides evidence on the potential for school-based interventions to influence socioemotional development during this formative stage.

More broadly, our research also contributes to the literature on large-scale, educational reforms. Many effective educational interventions have relied on non-governmental inputs, including NGOs, or by "outsourcing" the management of schools through private-public partnerships (Banerjee et al., 2017; Eble et al., 2021; Romero et al., 2020). By contrast, efforts to improve government-provided education using only public sector resources have often struggled to enhance student learning (de Barros et al., 2024).[4] In this study, we examine a government-implemented intervention that operates solely with existing school personnel and shed light on the role of socioemotional skill production within that effort. This approach offers rare evidence on how governments can implement large-scale reforms to boost public sector efficiency.

1.2 Research questions

The study's population of interest consists of Moroccan public lower secondary school students. We focus on students in the first two grades of lower secondary school (grades

[2]Training on socioemotional skills is also often included in (yet largely not separable from other components of) school-based entrepreneurship training, financial literacy training, and public health interventions for secondary-school students (e.g., Chioda et al., 2021).

[3]In a parallel, ongoing randomized evaluation, Kevin Carney and Avinash Moorthy are currently evaluating the effects of a "happiness curriculum" in public schools in New Delhi, India.

[4]Private-public partnerships and government collaborations with NGOs are not at all immune to the challenges of scaling up promising interventions within the government system; for an unsuccessful attempt to scale a well-known early-childhood intervention within India's childcare system, see Arteaga et al. (2024).

7 and 8) and investigate their development over one school year. We seek to answer the following research questions concerning main outcomes and final outcomes.

1. **Main outcomes.** Does a school's assignment to the socioemotional support intervention impact students' intra- and interpersonal skills?

2. **Final outcomes.** Does a school's assignment to the socioemotional support intervention impact students' academic skills and dropout?

The study seeks to answer the following secondary research questions.

1. Does a school's assignment to the socioemotional support intervention impact the components and subcomponents of the above-mentioned families of main and final outcomes?[5]

2. Does a school's assignment to the socioemotional support intervention impact the above-mentioned families of main and final outcomes for two subgroups of students (students flagged for being at risk of dropping out and female students)?

To better understand mechanisms and impacts on potential mediators, the study will also explore the following ancillary research questions.

1. Does a school's assignment to the socioemotional support intervention impact students' well-being?

2. Does a school's assignment to the socioemotional support intervention impact students' study habits outside of school and students' participation in extracurricular activities?

3. Does a school's assignment to the socioemotional support intervention impact students' creativity?

Finally, to better understand take-up and implementation fidelity, the study will also investigate the extent to which the socioemotional support intervention is implemented well and taken up as intended. In doing so, we will pay particular attention to whether students know their school's social specialist, have met with the specialist, and have attended a workshop with the specialist.

[5]For example, the "family" of academic skills includes four subject-wise components (Arabic, French, mathematics, science), which can be further subdivided into sub-components (e.g., two indices of math items capturing at-grade-level content vs. below-grade-level content, respectively).

2 Research design

2.1 Basic methodological framework

This is a cluster-randomized controlled trial with a waitlist design. We randomly assigned 84 lower secondary schools to receive the socioemotional support intervention immediately ("treatment group") and 116 schools to potentially receive it after the study's completion ("control group"). Random assignment of schools to the socioemotional support intervention allows us to study the causal effect of being assigned to this intervention (the intent-to-treat, or "ITT" effect) on the outcomes of interest among students.

2.2 Hypotheses

Our hypotheses and hypothesis tests follow the research question outlined in Section 1.2. We investigate these research hypotheses following a pre-specified, ordered hierarchy of tests, accounting for multiple hypothesis testing. We detail these hypothesis tests, as well as their order and branching logic in Section 3.3.

Our primary hypotheses are that a school's assignment to the socioemotional support intervention impacts students' intra- and interpersonal skills (main outcomes), and that a school's assignment to the socioemotional support intervention impacts students' academic skills and dropout (final outcomes).[6] Our secondary research hypotheses posit impacts on these main and final outcomes among two subgroups of students: students at risk of dropping out and girls.

Using machine learning-based methods, we will also explore the potential heterogeneity of effects on the study's main and final outcomes across a large vector of student and school background characteristics (following Carlana et al., 2022; Athey and Imbens, 2016). This exploratory analysis will include the subgroup effects among the students not explicitly pre-registered above (i.e., students not flagged as being at risk of dropping out and boys).

2.3 Outcome variables

This study is based on primary data collected in two data collection rounds, additional data on implementation quality and take-up, and administrative data. Primary data

[6] Downstream effects on academic skills may be positive or negative and we do not hypothesize a direction of these effects. While socioemotional skills may complement the production of academic skills (leading to positive effects on academic skills), we may also observe multitasking, whereby emphasizing one type of skill can crowd out the other (leading to negative effects on academic skills). Our study's findings will shed light on whether either type of effect is at play.

collection includes a "baseline" and "endline" in all 200 schools sampled for the study (September 2024; June 2025). During each of these rounds, all students of our study sample complete group-administered, paper-based assessments for all four subjects. Those students sampled for Arabic and French also complete one-on-one interviews, capturing their oral language skills and all other non-test measures described below (including those related to socioemotional skills).

To aggregate responses to test and interview questions ("items"), we employ different types of item response theory (IRT) models, each selected to align with the structure and characteristics of a given instrument.[7] For survey instruments with a common set of ordinal response categories, we use the graded response model (GRM). For instruments with varying ordinal response categories, we use the partial credit model (PCM). For instruments with strictly binary items (including test questions rated as correct vs. incorrect), we use the two-parameter logistic (2PL) model. Throughout, we use overlapping items (or "anchors") to map test scores onto common scales across grades and assessment rounds.[8] For select measures (indicated below), we further aggregate IRT scores into indices by calculating the inverse covariance-weighted average across these scores [9]. Each score (or index) will be standardized by subtracting the mean and dividing by the standard deviation of the distribution of the scores of the students in the control group.[10]

2.3.1 Main outcomes

Interpersonal skills. Our global indicator of interpersonal skills consists of a social skills index that combines measures of students' pro-sociality and emotion perception. It integrates the pro-sociality subscale of the Strengths and Difficulties Questionnaire (SDQ) and the Perceiving AI Generated Emotions scale (PAGE). The SDQ pro-social component, self-rated for 11- to 17-year-olds, measures students' tendencies to cooperate, help, and engage positively with peers. Pro-social behaviors are fundamental to strong peer relationships and have been linked to higher levels of school engagement and lower behavioral problems (Goodman, 1997). The PAGE scale, a 16-item assessment of emotion perception, evaluates students' ability to recognize and interpret emotions in facial expressions, a critical component of emotional intelligence (Weidmann and Xu, 2024).

[7] See Jacob and Rothstein (2016) for an accessible introduction to Item Response Theory in the economics literature.

[8] Across all models, items with negative discrimination parameters are systematically excluded from the analyses, as these items do not contribute to the accurate differentiation of respondents based on their abilities.

[9] Appendix 3.3 briefly describes each of the models used to aggregate our items, along with the methods used to further aggregate IRT scores into indices

[10] In this document, we standardize these scores as per the control-group distribution at baseline. Once we have collected the endline data, we will standardize these scores as per the control-group distribution at endline.

Intrapersonal skills. Our global indicator of intrapersonal skills combines two indices by calculating their inverse covariance-weighted average: a perceived control index, which includes growth mindset, locus of control, and self-efficacy, and a self-discipline index, which measures self-regulation, diligence, and work discipline.

The perceived control index captures students' beliefs about their agency and ability to influence their academic and personal success. It covers three well-established constructs: growth mindset, locus of control, and self-efficacy. Together, using these three components, we assess students' perceived control—beliefs about one's ability to influence outcomes, which have been shown to be fundamental drivers of motivation and behavior. The growth mindset scale assesses students' beliefs about intelligence as a malleable trait. The locus of control measure, adapted from (Huillery et al., 2025), assesses whether students attribute outcomes to their own actions or to external forces. The self-efficacy scale, derived from the PISA self-efficacy battery, measures the confidence of students in overcoming academic challenges and persevering through difficulties.

The self-discipline index captures students' ability to regulate their behavior, persist through challenges, and maintain focus on tasks, integrating measures of diligence, discipline, and self-regulation. The Short Grit Scale (Duckworth and Quinn, 2009)), measures perseverance and sustained effort, assessing students' willingness to work through setbacks and delays in gratification. To evaluate work discipline, we followed the instrument used by Huillery et al. (2025), which evaluates the ability of students to stay on task, complete assignments, and maintain attention, reflecting their self-management skills in academic settings. Finally, the self-regulation measure, based on the Short Self-Control Scale (8-item version), assesses impulse control and students' ability to manage distractions and emotional responses.

2.3.2 Final outcomes

Academic skills. Our study includes four subsamples of students; one for each academic subject (see section 2.7). In our analyses of effects on overall academic skills, before reporting on subject-wise results, we will stack the four subsamples and use their respective standardized test scores as the outcome (Arabic, French, mathematics, or science).

We measure students' mathematical skills across five content domains: numbers, geometry, algebra, data and probability, and measurement. These domains, aligned with the TIMSS and PISA frameworks, offer a well-rounded evaluation of students' mathematical abilities. The assessment also incorporates three cognitive dimensions—knowing, applying, and reasoning—to measure not only students' ability to recall and recognize mathematical

concepts but also their capacity to apply mathematical procedures and engage in higher-order thinking.

We measure students' science skills across life sciences, physical sciences, and earth sciences, covering content areas such as biology, chemistry, physics, and environmental science. These domains align with the TIMSS and PISA frameworks, ensuring that the assessment reflects internationally recognized benchmarks for scientific literacy. Like the mathematics assessment, the science evaluation incorporates three cognitive dimensions—knowing, applying, and reasoning—to capture students' ability to recall key scientific concepts, apply them in real-world contexts, and engage in analytical reasoning and problem-solving.

The paper-based component of the Arabic and French language assessment captures reading comprehension and written production. The reading component required students to extract explicit information, draw logical conclusions, synthesize ideas, and critically analyze texts, while the writing component assessed their ability to formulate clear, structured, and coherent written responses. Both components align with the PIRLS and PISA frameworks.

The oral language assessment complements the written component by evaluating listening comprehension, oral reading fluency, and speaking skills. Listening comprehension tasks measure students' ability to understand spoken language through contextually relevant passages, requiring them to process and interpret information. Oral reading tasks assess fluency, accuracy, and expression while speaking tasks evaluate vocabulary usage, verbal clarity, and the ability to articulate ideas effectively.

Dropout. We will observe whether a student who participated in the study's baseline assessment has left the public school system at the end of the current school year as well as the beginning of the following school year. To this end, we will use official enrollment records from the Ministry.

2.3.3 Additional outcomes

Well-being. We construct an index of student well-being, integrating measures of belonging, bullying, and perceived stress administered through one-on-one, paper-based student interviews. The belonging scale, adapted from PISA's school climate module, evaluates students' sense of connection and inclusion in their school environment. The bullying scale, also drawn from PISA, captures experiences of peer victimization, with scores reversed to align with positive well-being outcomes. Finally, the PSS-4, a widely

used psychological scale developed by Cohen et al. (1983), measures students' perceived stress and ability to manage challenges.

Creative thinking skills. We measure fluency, originality, and elaboration in creative tasks using the Torrance Tests of Creative Thinking (TTCT) (Torrance, 1968, 1998) instrument administered through one-on-one interviews. The TTCT is a widely recognized assessment designed to capture students' ability to generate ideas, think flexibly, and expand on initial concepts.

Study habits. We collect data on students' habits outside of school. In particular, we collect self-reported measures about students' extracurricular activities and the time spent on homework after school. We focus on two binary variables (borrowed from PISA) that indicate whether, for a typical school week, a given student reports spending at least 30 minutes a day on homework or extracurricular activities after school, respectively.

2.3.4 Non-outcome measures

Take-up and implementation quality. We will assess the intervention's implementation fidelity and take-up through one round of process-monitoring school visits, where we will evaluate whether the socioemotional support intervention is delivered as intended. During these visits, we will observe and document the extent to which implementers adhere to the planned intervention structure, ensuring that some key activities and instructional strategies have been executed as designed.[11] Additionally, during the one-on-one child interviews conducted at endline, we will also collect data on whether students (in both treatment and control groups) know their school's social specialist, whether they have met with the specialist, and whether they have attended a workshop with the specialist.[12] We will also estimate the intervention's impacts on students' familiarity and interactions with their school's social specialist.

Social desirability. To be able to test (and account) for the potential presence of experimenter demand effects, we also measure students' propensity to give socially desirable responses (following Dhar et al., 2022). To this end, we administer a short

[11] Unfortunately, due to scheduling concerns, it is unlikely that we will be able to conduct unannounced, "surprise-visit" observations of the workshops.

[12] During the one-on-one student interviews at endline, we will ask students whether they have participated in a workshop with the school's social specialist since returning to school from the Aïd el-Fitr holiday at the end of Ramadan. This holiday provides a clear marker that students can easily remember, rendering an approximately two-month-long recall period.

survey module during one-on-one interviews and construct the Marlowe-Crowne social desirability scale (Crowne and Marlowe, 1960). The survey module asks respondents if they have several "too-good-to-be-true traits" (such as always being a good listener); those who report more of these traits are scored as having a higher propensity to give socially desirable responses.

Other covariates. We have access to the Ministry's school-level administrative data for the universe of public lower secondary schools in the country (for the years 2021, 2022, and 2023). This data provides us with rich background information (including on staffing, enrollment, and school infrastructure, for example). We also have access to student-level, administrative information for the students in our study schools (including on their academic performance at the time they graduated from primary school).

2.4 Context and intervention

2.4.1 Context

Public provision of education, as governed by the Ministry of Education (MENFPESRS or Ministère de l'Education Nationale, du Préscolaire et des Sports), is the most common type of primary and secondary education in Morocco. Even though the share of private enrollment has increased over the recent years, as of 2019, public schools still served 83 percent of primary school students and 89% of lower secondary school students in the country (The World Bank, 2025). Enrollment rates are high, with net enrollment rates of 99.6 percent at the primary level and 90.6 percent at the lower secondary level in 2019 (The World Bank, 2025).

From 2018 to 2023, the lower secondary completion rate increased from 64.5 to 74.2 percent (from World Bank World Development Indicators). However, these high enrollment and completion rates mask very low student performance. In the 2021 PIRLS reading assessment, Morocco's fourth-graders ranked second to last (out of 57 countries), and more than half of the students (59 percent) did not reach the minimum proficiency benchmark (Mullis et al., 2023). In the 2023 TIMSS math assessment, Morocco's fourth graders ranked third to last (out of 58 countries), and more than half (54 percent) did not attain the minimum proficiency benchmark (von Davier et al., 2024). At the secondary level, in the 2023 TIMSS math assessment Morocco's eighth graders ranked second to last out of 42 countries, and 64% of grade 8 students were below the lowest international benchmark in mathematics (von Davier et al., 2024).

We conduct this study within the context of a broader reform effort that seeks to address these low learning levels in Morocco's public lower secondary schools. The reform—which is locally known as the "Pioneer School Program (PSP)"—is inspired by the Global Education Evidence Advisory Panel (GEEAP, 2023). The socioemotional skill intervention is one intervention component of that reform. The program was launched in 232 of Morocco's public lower secondary schools in September of 2024.[13] The Ministry considers the program to be its flagship intervention in lower secondary schools, and it intends to scale the program to another 500 schools in the 2025-26 school year, covering approximately 32 percent of the students in the country. The remaining schools are expected to be reached by 2028.

2.4.2 The socioemotional support intervention

The socioemotional support intervention was inspired by the socio-ecological model (Kilanowski, 2017), which linked school dropout to several levels: individual, relational, community, and societal. In practice, the intervention consists of training and equipping social specialists to give four workshops per year in small classes for students in the first and second year of lower secondary education (grades 7 and 8). These workshops aim to promote awareness of the role of the social specialist, build a relationship of trust with the students, and develop intrapersonal and interpersonal skills to strengthen resilience (self-management capacity). The workshops are also designed to dedicate particular attention to at-risk students in need of individualized support. Beyond these workshops, the intervention intends to promote awareness of the social specialists working in the schools and build relationships of trust between the specialists and the students.

The staff member in charge of implementing the socioemotional support intervention is the school's social specialist.[14] Six social specialists across Morocco were selected to receive training and thereafter train other social specialists in the schools we assigned to receive the intervention. Social specialists are provided with training slides, and students are provided with booklets for each workshop as well as a notebook to write down their feelings and emotions.

Each workshop focuses on a specific topic related to intra- or interpersonal skills. The first workshop is about self-awareness, focusing on self-assessment (qualities, weaknesses) and

[13]The 232 schools volunteered to participate in the first year of the reform. The launch in lower secondary schools comes on the heels of launching the PSP program in primary schools one year prior. We evaluate the overall effect of (the bundle of interventions provided by) the reform in both types of schools in other research.

[14]The social specialist is a civil servant who has been trained after a competitive examination open to graduates in psychology, philosophy, or social science. Their work consists of ensuring student well-being at school and identifying students in need of support or attention. A social specialist is assigned to a school and works on-site a total of 24 hours per week.

development mindset. The second workshop is about students' quality of interpersonal relationships and social skills, focusing on identifying and understanding emotions, and regulating emotions; the third workshop is about self-regulation and emotions, focusing on collaboration with peers and students' sense of belonging; and the fourth workshop is about cooperation and collaboration, focusing on active listening, empathy, and conflict resolution.

2.5 Sampling of schools

Our sample of schools includes 200 lower-secondary schools. We constructed the sample in two steps, using administrative data on the universe of government lower secondary schools, which includes the 232 lower-secondary PSP schools in Morocco. First, using the "post-double-selection" (PDS) methodology (Belloni et al., 2012, 2011, 2014, 2016), 30 variables were identified that were predictive of either test scores or participation in the reform. Second, using these 30 selected variables, Mahalanobis nearest-neighbor matching (without replacement) was used to identify 100 pairs of PSP schools that were very similar to one another (in terms of their Mahalanobis distance).[15]

Table 1 provides an analysis of the sample's representativeness, both in relation to the population of lower secondary schools in Morocco and in relation to other Pioneer Schools in the country, alongside balance checks. Columns (1) to (3) compare the 200 lower secondary schools in the study's sample of schools with the 2,344 remaining (public) lower secondary schools in Morocco[16]. Columns (4) to (6) present the representativeness within the reform program by comparing the 200 programs (PSP) lower secondary schools in the study sample with the 32 other PSP lower secondary schools in Morocco. Overall, schools in the study sample have slightly more teachers, fewer rural locations, have more students, and serve students with a slightly higher average primary and middle-school school leaving exam score. Columns (7) to (9) present the balance among the two groups of schools included in this study, which we discuss below in section 2.11. Importantly, as per the overall F-test, we do not find evidence of systematic imbalance across the treatment and control groups, which suggests our randomization was successful in creating two groups of schools that are, in expectation, indistinguishable from each other (Column 9).[17]

[15]We initially considered assigning 100 schools to the intervention, in a pairwise, stratified trial, but we discarded this plan later on. The following subsection describes our random assignment of 84 schools to the intervention.

[16]At the stage of finding matched control schools for each of the pioneer schools, we had received data on 2,544 public lower secondary schools

[17]While these balance checks focus on school characteristics, below, we also present additional balance checks at the student-level.

Our power calculations account for the uneven ratio of schools assigned to the treatment and control conditions, as well as the same 9-percent attrition rate we observed in a separate study conducted by two of us in Morocco's public primary schools last year. We assume an intra-cluster correlation (ICC) of 0.1, R-squared of 0.5, power of 0.8, and alpha of 0.05.[18] For the overall sample of students, these calculations indicate the study is well-powered to detect even small effects. After (conservatively) applying Šidák corrections to account for multiple hypothesis testing, the minimal detectable effect (MDE) is 0.107 standard deviations for academic skills and 0.125 standard deviations for intra- and inter-personal skills. For the pre-specified subgroup analyses concerning at-risk students, these MDEs are 0.163 and 0.208, respectively.

For the overall sample of students, these Šidák-corrected MDEs are well in line with the intent-to-treat effects on test scores commonly found in evaluations of large-scale education programs in less-developed countries. They are just above the median effect on student learning of 0.10 standard deviations reported by Evans and Yuan (2022), and they are well below the MDEs common for interventions targeting socio-emotional skill outcomes reported by Wilson et al. (2025). As expected for additional subgroup analyses, in light of the required adjustments for multiple hypothesis testing and the smaller sample of students qualifying as at-risk, the corresponding estimates for at-risk students will be noisier.[19]

2.6 Random assignment of schools

Out of the 200 PSP schools, we randomly assigned 84 schools to receive the socioemotional support intervention (and the remaining 116 schools to not receive that intervention). Specifically, to conduct a stratified randomization strategy, we started by calculating a principal component based on the average primary school passing grade, number of teachers, number of students, and fraction of female teachers in the 200 schools. We then used this principal component to sort the schools into an alternating pattern of triplets and pairs (84 groups in total). Finally, within each group, we randomly assigned one school to receive the socioemotional support intervention.

[18] Stacking the baseline test scores across all subjects, the intra-school correlation of baseline learning *levels* is 0.1. Yet, we believe the intra-cluster correlation of *growth* in student learning between the baseline and endline assessments to be lower than the ICC of baseline learning levels. Accordingly, our assumption for the ICC of endline test score residuals should be a slightly conservative overestimate.

[19] At the same time, given the greater need for support among at-risk students, effects may be larger in this subgroup.

2.7 Sub-sampling of students

The study's unit of analysis is a student. In each of the study's lower secondary schools, surveyors were given a "priority" list of randomly selected students for each grade. We constructed these lists before the baseline data collection began, using enrollment records. The lists allowed for random replacements if students were absent on the day of the baseline assessment.

Students were sampled to take the assessments in one subject, only (Arabic, French, mathematics, and science). For the written Arabic and French assessments, as well as the one-on-one oral assessments and interviews capturing socioemotional skills, we subsampled up to 12 students per school (6 students from grade 7, and 6 students from grade 8).[20] For the written math and science assessments, we subsampled up to 18 students per school (all attending grade 7).[21]

Specifically, the students in each grade and track (APIC or ASCG) were randomly split into four groups, one for each subject.[22] Within each group of students sampled for a subject, students were randomly ordered into a ranked list of students. This randomly ordered list of students ensured that each grade and subject had the required number of students sampled for the survey, as well as that the number of students on the list from each track was proportional to the number of enrolled students in each track for that subject. Thus, a randomized priority list of students was generated, which had students with "high" priority who enumerators would try to assess/survey first before moving down the list to the students with "low" priority.

The study's effective sample consists of 11,140 students (6,488 in the control group and 4,652 in the treatment group). These are all the students who took the baseline assessment and are now being tracked to the endline assessment. Among these students, 2,140 were tested in Arabic, 2,150 in French, 3,429 in math, and 3,421 in science.

2.8 Theory of Change

In terms of *need*, the intervention seeks to address two related challenges common to many low- and middle-income countries. First, our qualitative insights into Morocco's lower secondary schools describe an environment that is largely unconducive for developing

[20] We also sampled and tested up to 6 students in grade 9, but since the socioemotional skill intervention focuses on the lower grades only, we exclude them from our data.

[21] For these two subjects, we did not include students from grades 8 or 9 as the broader reform did not yet include math and science interventions for these grades.

[22] APIC refers to "Année Secondaire Collégial Originel Parcours International" and APIC refers to "Année Secondaire Collégial Général".

socioemotional skills, with high prevalence levels of bullying and intimidation, conflicted relationships between students and teachers, feelings of insecurity, and general distress and discomfort among students. Second, while personnel exists (in this case, social specialists), existing services that could promote socioemotional skills appear ineffective, with limited support for students, support relying solely on voluntary participation, social specialists being assigned to administrative tasks (or only to the most serious, disciplinary cases), a lack of clarity about responsibilities, and absence of tools and training. Taken together, the study context thus appears to reveal serious levels of need for the intervention and a high likelihood that the intervention will generate a meaningful contrast with the status quo.

The *inputs* offered by the intervention are trainings for social specialists working in the schools and guidebooks on how to implement the intervention, following a government order for the specialists to do so (for a description of the intervention, see section 2.4). Other inputs for schools selected for the ongoing reform are held constant; the intervention is implemented with the existing resources, with no additional staff or pay allocated to intervention schools.

The expected *outputs* are that social specialists hold four workshops per year in small classes for students in the first and second years of lower secondary education and that students participate in these workshops.

The expected *intermediate outcomes*, mechanisms, and secondary outcomes are that students' well-being increases, their study habits improve, and students think more creatively. The intervention may also trigger changes in students' participation in extra-curricular activities, but to the extent that these activities do not foster academic or socio-emotional skills, this effect may have detrimental consequences on downstream variables.

The expected *outcomes* are that students improve their interpersonal (or social) skills and improve their intrapersonal skills. The former relates to positive impacts on personal perception and pro-sociality. The latter relates to positive impacts on students' perceived control (incl. growth mindset, locus of control, and perceived self-efficacy) and students' self-regulation and discipline (incl. self-control and grit).

Lastly, the expected *impact* is that students improve their academic skills (in Arabic, French, math, and science) and become less likely to drop out of school.

2.9 Variations from the intended sample, and non-compliance

We base the study's expected attrition rate on a similar study conducted just one year prior, in Morocco's public primary schools. We do not expect attrition to exceed the 9 percent that we have thus factored into our statistical power calculations (see section 2.5).

We may encounter differential attrition at endline if students in the treatment group have changed their propensity to come to school. This may happen if the intervention affects student dropout. We discuss how we will address differential attrition in section 3.2.2.

We believe cross-over or "contamination" across experimental groups is highly unlikely. We randomized schools in the country to receive (or not receive) the socioemotional intervention, which limits contamination from one school to another. We also closely worked with the Ministry that oversees implementation and monitoring. Moreover, we do not believe students will switch their enrollment from a control school to a treatment school, because of the intervention.[23]

2.10 Data collection and processing

Primary data collection is primarily handled by Ministry staff external to the study schools, with the support and oversight of MEL/J-PAL staff. We adhere to strict data collection protocols, including spot-checks and accompaniments, and weekly monitoring and debriefs for enumerators (see Glennerster, 2017; J-PAL, 2017).

2.11 Initial findings from the study's baseline

Table 1 and 2 examine balance in school characteristics, learning outcomes, socioemotional skills, and student characteristics between the comparison schools and the two types of PSP lower secondary schools (with or without the socioemotional support intervention). As expected, given the random assignment of the socioemotional support intervention, there are no significant differences in school characteristics between the 116 PSP schools without the intervention and the 84 PSP schools with the intervention. Furthermore, the comparison schools are generally similar to both types of PSP schools, and joint F-tests confirm no systematic differences across these three groups.

Table 3 provides psychometric properties of the academic skill assessments and measures of socioemotional skills, including properties based on Classical Test Theory (CTT; columns 4-7) and Item Response Theory (IRT; columns 8-10).[24]

In Arabic, math, and science, students attempted almost all test questions, leaving only 4.1, 2.7, and 4.3 percent of the questions unanswered (see column 5). For French, this percentage is only slightly higher (14.5 percent and 7.5 percent, depending on students'

[23] At the end of the study period, we will have data on student transfers, which will allow us to verify this claim.

[24] Since the measures in Panel B largely use Likert-type answer formats, we do not report CTT-based properties for them.

grade). Taken together, this finding suggests that the assessments were of manageable duration and produced limited respondent fatigue among students.

In addition, almost all test items performed well. Hardly any test items had to be removed due to unfavorable measurement properties (see column 2), and the average test item discriminated very well (see column 8)[25]. In addition, the tests proved to be internally consistent, with average item-test correlations ranging from 0.35 and 0.52.

The average conditional reliability, reported in column (10) of Table 3, measures the precision of each instrument across the spectrum of respondent abilities.[26] For all of our academic skill measures, we find reliability values close to 0.8 (or higher), which indicates that the instruments consistently measure the constructs of interest with very high levels of precision. This is true for the full spectrum of student ability—while the tests were hard for students, and even though precision could be even further improved by including easier test questions, we do not worry about floor effects.

In turn, for the measures of socioemotional skills, the reliability values vary by instrument. The self-discipline index and creativity index measured students' skills with high levels of precision, with reliability values close to 0.8. In contrast, the remaining measures were noisier. Based on these findings from the study's baseline, we have already taken action to improve the precision of these instruments (e.g., by lengthening the surveyor training modules for these instruments).

Finally, Table 4 explores the relationship between students' academic skills and their socioemotional skills. Across almost all dimensions—and despite the initial noisiness of some of the measures at baseline—socioemotional skills are positively correlated with Arabic and French skills. The only exception is time spent on after-school activities, which exhibits a slightly negative correlation with Arabic skills and no correlation with French skills. This pattern may suggest that after-school activities reduce study time, thereby lowering academic performance. Importantly, these correlations remain robust even after controlling for students' propensity to provide socially desirable responses.

[25]Generally, a value above 0.5 or 1.0 is considered high, with the scale usually ranging from 0 to 2.0 depending on the specific IRT model being used.

[26]To report on the reliability of our measures, we calculate the average conditional reliability as

$$\frac{1}{N}\sum_{i=1}^{N}\left[1-\frac{\mathrm{SE}(\hat{\theta}_i)^2}{\mathrm{Var}(\hat{\theta}_i)}\right],$$

where N is the sample size, $\mathrm{SE}(\hat{\theta}_i)$ is the predicted standard error of the ability estimate for individual i, and $\mathrm{Var}(\hat{\theta}_i)$ is the predicted ability variance. Unlike marginal reliability, which uses the theoretical θ distribution, this measure reflects the predicted test performance for our specific sample. Cronbach's alpha (reported in column 6) is the corresponding Classical Test Theory-based measure of reliability. As expected, the two measures of reliability match each other very closely.

3 Empirical analysis

3.1 Statistical model of the effect of the socioemotional support intervention

Our identification strategy rests on the study's random assignment of schools to the two experimental groups. We will exploit this random assignment to estimate the causal effects of being assigned to the intervention through linear regressions. We will estimate the intent-to-treat effect of the socioemotional support component on the outcomes of interest using a regression strategy, comparing pioneer lower secondary schools that were randomly assigned to receive the intervention with pioneer lower secondary schools that were randomly assigned not to receive the intervention. For all outcomes, we use the following empirical specification:

$$Y_{igsr}^{t} = \alpha_r + \beta_1 T_{sr} + \boldsymbol{\delta'} \boldsymbol{X}_{igsr}^{t=0} + \epsilon_{igsr}^{t} \tag{1}$$

Here, Y_{igsr}^{t} is the outcome Y in period t for student i in grade g in school s and randomization strata r. In our primary analyses, Y_{isr}^{t} represents test scores. The α_r terms are strata fixed effects, T_{sr} is the treatment dummy indicating a school's random assignment to the socioemotional support intervention, and ϵ_{igsr}^{t} is the residual. To increase precision, all specifications include $\boldsymbol{X}_{igsr}^{t=0}$ as covariates. Measured at baseline ($t = 0$), $\boldsymbol{X}_{igsr}^{t=0}$ is a vector of baseline controls selected by a post-double selection (PDS) Lasso procedure on student and school characteristics, partialing out the strata fixed effects and, when available, $Y_{igsr}^{t=0}$ (a student's outcome of interest at baseline).[27]

The coefficient of interest is β_1, which captures the intent-to-treat effect of assignment to the socioemotional support intervention among schools in the Pioneer Schools Program.

In our analyses of treatment effects among student subgroups, we will interact the treatment indicator with the respective subgroup indicators. In our exploratory analyses, we will employ a machine learning-based method (causal forests) to investigate heterogeneous effects among subgroups of schools and students (following Carlana et al., 2022; Athey and Imbens, 2016).

[27] We thank Jacobus Cilliers for pointing us to Cilliers et al. (2024) and for recommending this approach. Following Cilliers et al. (2024), we will use the default, "plug-in" penalty parameter of Stata's *pdslasso* command (not cross-validation).

3.2 Statistical methods

3.2.1 Estimation

We will estimate equation (1) using ordinary least-squares (OLS) regressions. We will cluster standard errors at the school level (Abadie et al., 2022).

3.2.2 Rules for handling missing values

We expect to encounter two types of missing data: attrition (e.g., students not participating in the endline assessment) or missing values (e.g., students participating in the endline, but not answering specific questions therein).

Missing values due to attrition. We will address the first type of missing data as follows. First, we will document the overall attrition rate. Then, we will investigate whether attrition is systematically related to intervention assignment by fitting a version of equation (1) that replaces the outcome variable with an indicator variable for not participating in the endline. Next, if we find differential attrition, we will use inverse-probability weighting (IPW) and Lee (2009) bounds estimations to subject our findings to robustness tests.

Missing values due to non-response. We will address the second type of missing data as follows. For missing responses on outcome variables, we will scale responses using item-response theory (IRT) models that account for missing values by using concurrent calibration via marginal maximum likelihood estimation (Kolen and Brennan, 2004), given that non-response on specific questions is akin to missingness in any non-equivalent anchor test (NEAT) design in which not all respondents are administered the same questions. To maintain the largest possible sample size, we will exclude covariates with missing values. In robustness checks, we will also follow Zhao and Ding (2024), include these covariates, impute any missing values of baseline covariates with zero, and include a dummy variable that indicates the missingness of the corresponding baseline covariate.

3.2.3 Definition and rules for handling outliers

We do not expect to encounter outliers because all of our outcome variables are measured on pre-determined scales. Therefore, we will not seek to identify outliers or winsorize results.

3.3 Multiple hypothesis testing

Accounting for a pre-registered hierarchy of research hypotheses, we will adjust for multiple hypothesis testing by computing the sharpened false discovery rate adjusted q-values. Following Vivalt et al. (2024), we place our hypotheses into tiers (denoted K0, K1, and K2), which correspond to our prioritization of tests.

K0: Family of main outcomes: Intra- and interpersonal skills. We will compute q-values for these two variables (sets of skills), which we will refer to as our two K0 items. These are our highest priority outcomes.

K1: Family of final outcomes: Academic skills and dropout. These are our second-highest priority outcomes. We will compute the q-values for these two variables (outcomes), which we will refer to as our two K1 items, in combination with our two K0 items.

K2: Family of additional outcomes: Creativity, well-being, and study habits. These are our third-highest priority outcomes. We will compute the q-values for these three variables (outcomes), which we will refer to as our three K2 items, in combination with our four K0 and K1 items.

In addition to these three tiers of family-level tests, we also explore effects at the underlying component- and sub-component-levels. For example, the "family" of intrapersonal skills consists of two "components": an index of perceived control, and an index of self-discipline. In turn, the perceived control component consists of three sub-components: a measure of growth mindset, a measure of locus of control, and a measure of perceived self-efficacy. To provide another example, the "family" of academic skills includes four subject-wise components (Arabic, French, mathematics, science), which can be further subdivided into sub-components (e.g., two indices of math items capturing at-grade-level-content vs. below-grade-level content, respectively). The q-values for component-level estimates will be computed using the family-level items up to the given family level (K0, K1, or K2), in addition to the component-level items in the outcome's same family. For example, the q-values for effects on Arabic skills will account for the four family-level comparisons at the K0 and K1 levels, as well as the four component-level comparisons (Arabic, French, math, and science).

Within each family tier, and at each level (family, component, or sub-component), we will explore treatment effects among two subgroups: students identified as being at risk of dropping out, and girls. Across these two subgroups, we prioritize the tests for the subgroup of at-risk students. The q-values for these comparisons will be computed using all the items for that family, and up to the given level, in addition to the tests of

heterogeneous effects. For example, tests for effects on Arabic skills among students at risk of dropping out will account for the four comparisons at the K0 and K1 levels, the four component-level comparisons (Arabic, French, math, and science), as well as four additional subgroup tests (one per subject). Tests for effects on Arabic skills among girls will account for the four comparisons at the K0 and K1 levels, the four component-level comparisons (Arabic, French, math, and science), the four subgroup tests for students at risk of dropping out, as well as four additional subgroup tests (one per subject).

Our analyses of intervention take-up and implementation fidelity are largely based on descriptive statistics, and we will not account for these analyses in our adjustments for multiple hypothesis testing.

Table 1: *Sample of schools, representativeness, and balance tests*

	Representativeness (overall)			Representiativeness (PSP)			Balance Checks		
	Non-Study	Study	Difference	Other PSP	Study PSP	Difference	Control	Treatment	Difference
	(1)	(2)	(3)	(4)	(5)	(6)			
Number of teachers	29.61	32.38	2.77***	27.12	32.38	5.25**	32.88	31.69	-1.19
	[15.20]	[14.35]	(1.06)	[12.53]	[14.35]	(2.41)	[14.25]	[14.56]	(2.07)
Rural (%)	49.27	37.00	-12.27***	65.62	37.00	-28.62***	36.21	38.10	1.89
	[50.01]	[48.40]	(3.57)	[48.26]	[48.40]	(9.10)	[48.27]	[48.85]	(6.96)
Total Enrolment	774.30	913.21	138.90***	834.62	913.21	78.58	941.09	874.70	-66.38
	[457.18]	[466.13]	(34.22)	[457.39]	[466.13]	(86.48)	[487.00]	[435.60]	(65.59)
Female students (%)	47.30	47.09	-0.21	48.04	47.09	-0.95	47.46	46.58	-0.88**
	[4.90]	[2.90]	(0.23)	[3.53]	[2.90]	(0.65)	[2.99]	[2.72]	(0.41)
Age	14.60	13.94	-0.66***	14.58	13.94	-0.64***	13.96	13.91	-0.05
	[0.36]	[0.22]	(0.02)	[0.26]	[0.22]	(0.05)	[0.20]	[0.23]	(0.03)
Primary-school leaving exam score	6.85	6.94	0.09***	6.79	6.94	0.15***	6.94	6.94	0.00
	[0.28]	[0.28]	(0.02)	[0.23]	[0.28]	(0.04)	[0.26]	[0.31]	(0.04)
Middle-school leaving exam score	8.95	9.14	0.19*	8.75	9.14	0.39*	9.07	9.22	0.15
	[1.46]	[1.36]	(0.10)	[1.09]	[1.36]	(0.21)	[1.31]	[1.43]	(0.20)
Number of schools	2344	200	2544	32	200	232	116	84	200
Joint F-test (p-value)			0.00			0.00			0.13

Notes. This table reports on the study's sample of schools. Study refers to the 200 lower secondary schools included in the study's effective sample. Non-study refers to all other public lower secondary schools in Morocco. Study PSP refers to the 200 Pioneer Schools included in the study. Other PSP refers to the remaining 32 (lower secondary) Pioneer Schools in the country. Control refers to the 116 schools not assigned to receive the socioemotional support intervention, and Treatment refers to 84 schools assigned to receive this intervention. Difference reports on the regression-adjusted difference. Standard deviations are shown in brackets; standard errors are shown in parentheses. Standard errors are clustered at the school level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2: *Descriptive statistics and balancing checks*

	Number of observations		Balancing check		
	Control	Treatment	Control	Treatment	Difference
	(1)	(2)	(3)	(4)	(5)
Panel A: Learning outcomes					
All Stacked	6488	4652	-0.01	0.05	0.05
			[1.00]	[1.02]	(0.05)
Arabic	1257	883	-0.00	0.10	0.10
			[1.00]	[0.96]	(0.07)
French	1254	896	-0.01	0.03	0.04
			[1.00]	[1.06]	(0.08)
Math	2003	1426	-0.01	0.03	0.04
			[1.00]	[1.03]	(0.06)
Science	1974	1447	-0.00	0.04	0.05
			[1.00]	[1.01]	(0.06)
Joint F-test (p-value)					0.66
Panel B: Self-development and wellbeing					
Intra-personal skills	2511	1779	-0.02	-0.00	0.01
			[1.01]	[0.97]	(0.05)
Perceived Control	2511	1779	-0.02	0.01	0.02
			[1.02]	[0.98]	(0.05)
Self-discipline Index	2511	1779	-0.01	-0.01	-0.01
			[1.01]	[0.97]	(0.04)
Social index: Perceiving emotions (PAGE)	2511	1779	-0.00	0.00	0.01
			[1.00]	[1.01]	(0.05)
Well-being index	2511	1779	-0.01	0.00	0.01
			[1.00]	[0.99]	(0.05)
Creativity (Torrance)	2511	1779	0.00	0.03	0.02
			[1.00]	[1.00]	(0.07)
Spent more than 30 min/day doing homework after school	2511	1779	0.96	0.97	0.01
			[0.20]	[0.18]	(0.01)
Days a week spent on after-school activities	2511	1779	2.90	2.83	-0.07
			[1.75]	[1.76]	(0.07)
Social desirability	2383	1682	-0.01	0.01	0.02
			[1.01]	[0.95]	(0.05)
Joint F-test (p-value)					0.83
Panel C: Student characteristics					
Age	6488	4652	12.64	12.60	-0.03
			[1.06]	[1.04]	(0.03)
Primary school passing score	6488	4652	7.10	7.13	0.03
			[1.17]	[1.14]	(0.05)
Female (%)	6488	4652	48.81	48.95	0.13
			[49.99]	[49.99]	(1.12)
Joint F-test (p-value)					0.79

Notes. This table describes the study's sample of 200 schools and presents balance checks. Control refers to the 116 schools not assigned to receive the socioemotional support intervention, and Treatment refers to 84 schools assigned to receive this intervention. "Balancing check" reports on the regression-adjusted difference. Reversed outcomes were flipped, so higher scores represent desirable outcomes. Standard deviations are shown in brackets; standard errors are shown in parentheses. Standard errors are clustered at the school level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: *Psychometric properties*

	Number of items			Classical test theory (CTT)				Item response theory (IRT)		
	Effective	Excluded	Anchors	% correct	% NA	Cronbach's alpha	Mean item-test corr.	Mean discrimination	Mean difficulty	Avg. conditional reliability
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Assessments										
Arabic	47	1		44.40	4.10			1.26	0.25	0.88
French	53	0		31.69	10.93			1.37	1.15	0.89
Math	32	0		37.76	2.66	0.77	0.35	0.73	1.13	0.79
Science	32	0		43.13	4.27	0.76	0.34	0.75	0.97	0.80
Arabic: AC1	39	1		42.58	4.92	0.90	0.47	1.30	0.41	0.91
Arabic: AC2	31	1	23	46.16	3.32	0.85	0.44	1.04	0.11	0.85
French: AC1	37	0		34.73	14.53	0.93	0.52	1.56	0.75	0.91
French: AC2	31	0	24	37.17	7.51	0.92	0.49	1.26	1.36	0.88
Panel B: Other measures										
Intra-personal skills index	36	0								0.55
Perceived Control Index	11	0						0.34		0.51
Self-discipline index	0	0						0.42		0.78
Inter-personal skills: Social index	12	1						0.49		0.48
Wellbeing index	0	0						0.40		0.66
Creativity (Torrance)	7	0								0.79

Notes. Sample and unit of observation: 11,141 assessed students across the 200 schools in the study. This table reports on the measurement properties of student assessment instruments and survey instruments included in the study's baseline. "Anchors" refers to items also used on another instrument. "NA" refers to non-response. Mean discrimination and mean difficulty refer to the mean discrimination and difficulty parameters from a two-parameter logistic IRT model. Average conditional reliability represents the mean of individual-level reliabilities across the sample, based on each respondent's estimated ability and associated predicted standard error.

Table 4: *Bivariate associations between non-test score outcomes and student learning*

	Overall		By subject			
	Test score		Arabic		French	
	(1)	(2)	(3)	(4)	(5)	(6)
Intra-personal skills	0.27***	0.30***	0.24***	0.26***	0.28***	0.31***
	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)
Perceived Control Index	0.25***	0.25***	0.23***	0.23***	0.25***	0.26***
	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)
Self-discipline index	0.21***	0.23***	0.20***	0.21***	0.21***	0.24***
	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)
Inter-personal skills: Social index	0.23***	0.23***	0.29***	0.28***	0.16***	0.16***
	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)
Wellbeing index	0.20***	0.19***	0.20***	0.20***	0.17***	0.17***
	(0.02)	(0.02)	(0.03)	(0.03)	(0.02)	(0.03)
Creativity (Torrance)	0.20***	0.20***	0.19***	0.18***	0.20***	0.20***
	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)	(0.03)
Percentage spent more than 30 min/day doing homework after school	0.46***	0.44***	0.52***	0.50***	0.40***	0.37***
	(0.09)	(0.09)	(0.16)	(0.16)	(0.12)	(0.11)
Days a week spent on after-school activities	-0.03**	-0.03***	-0.04**	-0.04***	-0.02	-0.02
	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)

Notes. Sample and unit of observation: 6,488 assessed students across the 116 control pioneer schools not assigned to the intervention. Math and science scores are excluded as the above non-test score outcomes were not collected for students sampled for these two subjects. Columns 2, 4, and 6 control for the social desirability score. Reversed outcomes were flipped, so higher scores represent desirable outcomes. Standard errors are shown in parentheses. Standard errors are clustered at the school level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

References

Abadie, A., Athey, S., Imbens, G.W., Wooldridge, J.M., 2022. When Should You Adjust Standard Errors for Clustering? The Quarterly Journal of Economics 138, 1–35. doi:10.1093/qje/qjac038.

Alan, S., Boneva, T., Ertac, S., 2019. Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. The Quarterly Journal of Economics 134, 1121–1162.

Arteaga, I., de Barros, A., Ganimian, A.J., 2024. The Challenges of Scaling up Effective Child-Rearing Practices Using Technology in Developing Settings: Experimental Evidence From India. Journal of Research on Educational Effectiveness doi:https://doi.org/10.1080/19345747.2025.2450318.

Ashraf, N., Bau, N., Low, C., McGinn, K., 2020. Negotiating a Better Future: How Interpersonal Skills Facilitate Intergenerational Investment*. The Quarterly Journal of Economics 135, 1095–1151. doi:10.1093/qje/qjz039.

Athey, S., Imbens, G., 2016. Recursive Partitioning for Heterogeneous Causal Effects. Proceedings of the National Academy of Sciences 113, 7353–7360. doi:10.1073/pnas.1510489113.

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., Walton, M., 2017. From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application. Journal of Economic Perspectives 31, 73–102. doi:10.1257/jep.31.4.73.

Barrera-Osorio, F., de Barros, A., Filmer, D., 2024. Longterm Impacts of Primary School Scholarships: Evidence from Cambodia. Journal of Policy Analysis and Management 43, 10–38. doi:10.1002/pam.22533.

de Barros, A., Fajardo-Gonzalez, J., Glewwe, P., Sankar, A., 2024. The Limitations of Activity-Based Instruction to Improve the Productivity of Schooling. The Economic Journal 134, 959–984. doi:10.1093/ej/uead099.

Beaman, L., Herskowitz, S., Keleher, N., Magruder, J., 2021. Stay in the Game: A Randomized Controlled Trial of a Sports and Life Skills Program for Vulnerable Youth in Liberia. Economic Development and Cultural Change 70, 129–158. doi:10.1086/711651.

Belloni, A., Chen, D., Chernozhukov, V., Hansen, C., 2012. Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica 80, 2369–2429. doi:10.3982/ECTA9626.

Belloni, A., Chernozhukov, V., Hansen, C., 2011. Inference for high-dimensional sparse econometric models. doi:10.48550/arXiv.1201.0220. arXiv.

Belloni, A., Chernozhukov, V., Hansen, C., 2014. High-dimensional methods and inference on structural and treatment effects. Journal of Economic Perspectives 28, 29–50. doi:10.1257/jep.28.2.29.

Belloni, A., Chernozhukov, V., Hansen, C., Kozbur, D., 2016. Inference in high-dimensional panel models with an application to gun control. Journal of Business & Economic Statistics 34, 590–605. doi:10.1080/07350015.2015.1102733.

Carlana, M., La Ferrara, E., Pinotti, P., 2022. Goals and Gaps: Educational Careers of Immigrant Children. Econometrica 90, 1–29. doi:10.3982/ECTA17458.

Chioda, L., Contreras-Loya, D., Gertler, P., Carney, D., 2021. Making Entrepreneurs: Returns to Training Youth in Hard Versus Soft Business Skills. URL: https://www.nber.org/papers/w28845, doi:10.3386/w28845.

Cilliers, J., Elashmawy, N., McKenzie, D., 2024. Using Post-Double Selection Lasso in Field Experiments. Working Paper 10931. The World Bank. Washington, D.C. URL: https://openknowledge.worldbank.org/entities/publication/0cde089d-33ba-4f51-8c03-b25b5114d41a.

Cipriano, C., Strambler, M.J., Naples, L.H., Ha, C., Kirk, M., Wood, M., Sehgal, K., Zieher, A.K., Eveleigh, A., McCarthy, M., Funaro, M., Ponnock, A., Chow, J.C., Durlak, J., 2023. The state of evidence for social and emotional learning: A contemporary metaanalysis of universal schoolbased SEL interventions. Child Development 94, 1181–1204. doi:10.1111/cdev.13968.

Cohen, S., Kamarck, T., Mermelstein, R., 1983. A global measure of perceived stress. Journal of Health and Social Behavior 24, 385–396. doi:10.2307/2136404.

Crowne, D.P., Marlowe, D., 1960. A new scale of social desirability independent of psychopathology. Journal of Consulting Psychology 24, 349–354. doi:10.1037/h0047358.

von Davier, M., Kennedy, A., Reynolds, K., Fishbein, B., Khorramdel, L., Aldrich, C., Bookbinder, A., Bezirhan, U., Yin, L., 2024. TIMSS 2023 International Results in Mathematics and Science. Boston College, TIMSS & PIRLS International Study Center. URL: https://doi.org/10.6017/lse.tpisc.timss.rs6460, doi:10.6017/lse.tpisc.timss.rs6460.

Deming, D., 2017. The Growing Importance of Social Skills in the Labor Market. The Quarterly Journal of Economics 132, 1593–1640. doi:10.1093/qje/qjx022.

Dhar, D., Jain, T., Jayachandran, S., 2022. Reshaping Adolescents' Gender Attitudes: Evidence from a School-Based Experiment in India. American Economic Review 112, 899–927. doi:10.1257/aer.20201112.

Dillon, M.R., Kannan, H., Dean, J.T., Spelke, E.S., Duflo, E., 2017. Cognitive science in the field: A preschool intervention durably enhances intuitive but not formal mathematics. Science 357, 47–55. doi:10.1126/science.aal4724.

Dinarte-Diaz, L., Egana-delSol, P., Martínez Alvear, C., Rojas Alvarado, C., 2024. When emotion regulation matters: The efficacy of socio-emotional learning to address school-based violence in Central America. Working Paper IDB-WP-1585. IDB Working Paper Series. URL: https://www.econstor.eu/handle/10419/299415, doi:10.18235/0012854.

Duckworth, A.L., Quinn, P.D., 2009. Development and Validation of the Short Grit Scale (GritS). Journal of Personality Assessment 91, 166–174. doi:10.1080/00223890802634290.

Eble, A., Frost, C., Camara, A., Bouy, B., Bah, M., Sivaraman, M., Hsieh, P.T.J., Jayanty, C., Brady, T., Gawron, P., Vansteelandt, S., Boone, P., Elbourne, D., 2021. How much can we remedy very low learning levels in rural parts of low-income countries? Impact and generalizability of a multi-pronged para-teacher intervention from a cluster-randomized trial in the Gambia. Journal of Development Economics 148, 102539. doi:10.1016/j.jdeveco.2020.102539.

Edmonds, E., Feigenberg, B., Leight, J., 2023. Advancing the Agency of Adolescent Girls. Review of Economics and Statistics 105, 852–866. doi:10.1162/rest_a_01074.

Evans, D.K., Yuan, F., 2022. How Big Are Effect Sizes in International Education Studies? Educational Evaluation and Policy Analysis 44, 532–540. doi:10.3102/01623737221079646.

Ganimian, A.J., 2020. Growth-Mindset Interventions at Scale: Experimental Evidence From Argentina. Educational Evaluation and Policy Analysis 42, 417–438. doi:10.3102/0162373720938041.

GEEAP, 2023. Cost-Effective Approaches to Improve Global Learning: What Does Recent Evidence Tell Us Are Smart Buys for Improving Learning in Low- and Middle-Income Countries? Technical Report.

The World Bank. Washington, D.C. URL: https://thedocs.worldbank.org/en/doc/231d98251cf326922518be0cbe306fdc-0200022023/related/GEEAP-Report-Smart-Buys-2023-final.pdf.

Glennerster, R., 2017. The Practicalities of Running Randomized Evaluations: Partnerships, Measurement, Ethics, and Transparency, in: Banerjee, A.V., Duflo, E. (Eds.), Handbook of Economic Field Experiments. Elsevier. volume 1, pp. 175–243. doi:10.1016/bs.hefe.2016.10.002.

Goodman, R., 1997. The Strengths and Difficulties Questionnaire: A Research Note. Journal of Child Psychology and Psychiatry 38, 581–586. doi:10.1111/j.1469-7610.1997.tb01545.x.

Huillery, E., Bouguen, A., Charpentier, A., Algan, Y., Chevallier, C., 2025. The Role of Mindset in Education : A Large-Scale Field Experiment in Disadvantaged Schools. The Economic Journal , ueaf015doi:10.1093/ej/ueaf015.

J-PAL, 2017. J-PAL Research Protocols. URL: https://drive.google.com/file/d/0B97AuBEZpZ9zZDZZbV9abllqSFk/view.

Jackson, C.K., 2018. What Do Test Scores Miss? The Importance of Teacher Effects on NonTest Score Outcomes. Journal of Political Economy 126, 2072–2107. doi:10.1086/699018.

Jacob, B., Rothstein, J., 2016. The Measurement of Student Ability in Modern Assessment Systems. Journal of Economic Perspectives 30, 85–108. doi:10.1257/jep.30.3.85.

Kilanowski, J.F., 2017. Breadth of the socio-ecological model. Journal of Agromedicine 22, 295–297.

Kolen, M.J., Brennan, R.L., 2004. Test Equating, Scaling, and Linking. 3rd ed., Springer, New York, NY.

Krishnan, P., Krutikova, S., 2013. Non-cognitive skill formation in poor neighbourhoods of urban India. Labour Economics 24, 68–85. doi:10.1016/j.labeco.2013.06.004.

Lee, D.S., 2009. Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. The Review of Economic Studies 76, 1071–1102. doi:10.1111/j.1467-937X.2009.00536.x.

Mulhern, C., 2023. Beyond Teachers: Estimating Individual School Counselors' Effects on Educational Attainment. American Economic Review 113, 2846–2893. doi:10.1257/aer.20200847.

Mullis, I.V.S., von Davier, M., Foy, P., Fishbein, B., Reynolds, K.A., Wry, E., 2023. PIRLS 2021 International Results in Reading. Boston College, TIMSS & PIRLS International Study Center. URL: https://doi.org/10.6017/lse.tpisc.tr2103.kb5342, doi:10.6017/lse.tpisc.tr2103.kb5342.

Romero, M., Sandefur, J., Sandholtz, W.A., 2020. Outsourcing Education: Experimental Evidence from Liberia. American Economic Review 110, 364–400. doi:10.1257/aer.20181478.

Santos, I., Petroska-Beska, V., Carneiro, P., Eskreis-Winkler, L., Boudet, A.M.M., Berniell, I., Krekel, C., Arias, O., Duckworth, A.L., 2022. Can Grit Be Taught? Lessons from a Nationwide Field Experiment with Middle-School Students. Working Paper 15588. IZA Institute of Labor Economics. Bonn, Germany. URL: https://www.econstor.eu/handle/10419/265809.

Soland, J., Kuhfeld, M., Edwards, K., 2024. How survey scoring decisions can influence your studys results: A trip through the IRT looking glass. Psychological Methods 29, 1003–1024. doi:10.1037/met0000506.

The World Bank, 2025. Education Statistics (EdStats). URL: https://datatopics.worldbank.org/education/.

Torrance, E.P., 1968. Torrance Tests of Creative Thinking. Personnel, Princeton, NJ.

Torrance, E.P., 1998. Torrance Tests of Creative Thinking: Norms-technical manual: Figural (streamlined) forms A & B. Scholastic Testing Service, Bensenville, IL.

Vivalt, E., Rhodes, E., Bartik, A.W., Broockman, D.E., Krause, P., Miller, S., 2024. The Employment Effects of a Guaranteed Income: Experimental Evidence from Two U.S. States. URL: https://www.nber.org/papers/w32719, doi:10.3386/w32719.

Weidmann, B., Xu, Y., 2024. PAGE: A Modern Measure of Emotion Perception for Teamwork and Management Research. doi:10.48550/ARXIV.2410.03704. version Number: 1.

Wilson, S.J., Freeman, B., Hedberg, E.C., 2025. Empirical Benchmarks for Effect Size Interpretation and Study Planning with Social and Behavioral Outcomes. Journal of Research on Educational Effectiveness , 1–26doi:10.1080/19345747.2024.2427767.

Zhao, A., Ding, P., 2024. To Adjust or not to Adjust? Estimating the Average Treatment Effect in Randomized Experiments with Missing Covariates. Journal of the American Statistical Association 119, 450–460. doi:10.1080/01621459.2022.2123814.

Appendix A: Measurement

IRT models are used to estimate the probability of answering an item correctly as a function of a latent parameter representing the students ability and of parameters relating to the item (Jacob and Rothstein, 2016). To aggregate responses to test and interview questions ("items"), we estimate different types of item response theory (IRT) models, each selected to align with the structure and characteristics of a given instrument. The IRT models share two fundamental assumptions. First, local independence, which means that the probability of a correct response to an item depends only on ability, and item parameters. Second, they assume that there is a single latent dimension that explains item performance.

This appendix briefly describes each of the models used to aggregate our items, along with the methods used to further aggregate IRT scores into indices. To estimate students' ability scores, each of our IRT methods uses expected a posteriori (EAP) scoring. Soland et al. (2024) show that, for EAP scoring, failure to account for between-group variation in ability distributions may lead to biased estimates of treatment effects in educational interventions. To avoid such bias, we follow the guidance in Soland et al. (2024) and estimate scores using a two-group IRT model. Thus, the treatment and control groups are allowed to have separate ability distributions (i.e., different means and variances) while assuming common item parameters.

3.3.1 Two-parameter logistic (2PL)

The two-parameter logistic (2PL) model estimates the probability of answering an item correctly based on the examinees ability (θ) and each items difficulty (b_i). This type of model is only applicable to instruments with strictly binary items (including test questions rated as correct vs. incorrect), and, therefore, we apply it to the language, math, and science tests, along with the perceived emotions instrument (PAGE).

The two-parameter logistic (2PL) model is given by:

$$P(Y_{ij} = 1 \mid \theta_j) = \frac{1}{1 + \exp[-a_i(\theta_j - b_i)]} \tag{2}$$

Where θ_j is the latent ability of student j, a_i is the discrimination parameter, and b_i is the difficulty of item i.

3.3.2 Graded response model (GRM)

For Likert-type items with ordered categorical responses (e.g., 1 = Never to 5 = Always), we use the Graded Response Model (GRM), appropriate for instruments such as *SDQ (pro-sociality)*.

The probability of responding in category k or higher is given by:

$$P(Y_{ij} \geq k \mid \theta_j) = \frac{1}{1 + \exp[-a_i(\theta_j - b_{ik})]}$$

where b_{ik} represents the threshold between category $k-1$ and k. This model captures both item sensitivity (via a_i) and the varying difficulty of response thresholds.

3.3.3 Partial credit model (PCM)

For instruments with varying ordinal response categories, we use the partial credit model (PCM). This model provides partial credit by estimating item-specific step difficulties, appropriate for behavioral measures that score a range of performance levels. We use the Partial Credit Model (PCM) for instruments with items where score categories are not necessarily equally spaced or ordered, including: *Perceived Control*, *Diligence and Discipline*, *Well-Being*, and *Self-Discipline*.

$$P(Y_{ij} = k \mid \theta_j) = \frac{\exp\left(\sum_{m=0}^{k}(\theta_j - \beta_{im})\right)}{\sum_{l=0}^{K_i}\exp\left(\sum_{m=0}^{l}(\theta_j - \beta_{im})\right)}$$

where β_{im} is the difficulty of step m for item i, and K_i is the maximum score for that item. This model is more flexible than the GRM when step intervals vary substantially.

3.3.4 Poisson principal component analysis (PCA)

To construct the Creativity Index, we use a Poisson principal component analysis (Poisson PCA). Unlike standard PCA, which assumes continuous variables, Poisson PCA models the discrete, non-negative nature of count data more accurately. This approach aligns with the creativity measure, which captures the number of responses students produce.

3.3.5 Aggregation of latent scores

Latent scores $(\hat{\theta}_{j1}, \hat{\theta}_{j2}, \ldots, \hat{\theta}_{jn})$ are estimated using the empirical Bayes method, which computes scores based on students' observed item responses and the estimated item

parameters. To account for potential bias that can be introduced by shrinkage, we follow Soland et al. (2024) and estimate latent traits using a two-group IRT model. This model allows the treatment and control groups to have separate ability distributions while holding item parameters fixed.

To compare scores across baseline and endline assessments, we use a standard IRT linking procedure. Specifically, we estimate item parameters from the control group at endline and then hold those parameters fixed when scoring the baseline data. This approach ensures scores are placed on a common scale and changes over time are interpretable.

After estimating individual-level latent scores $(\hat{\theta}_{j1}, \hat{\theta}_{j2}, \ldots, \hat{\theta}_{jn})$ from each IRT model, we construct composite indices taking the inverse covariance-weighted average of component scores. Specifically, we compute the inverse covariance matrix of the component scores, normalize these weights to sum to one, and use them to compute a weighted average:

$$\text{Index}_j = \sum_{i=1}^{n} w_i \hat{\theta}_{ji}, \quad \text{where } w_i \propto \frac{1}{\text{Var}(\hat{\theta}_i)}$$