MEL
MOROCCO INNOVATION
AND EVALUATION LAB

UM6P
University
Mohammed VI
Polytechnic

Beyond Basics:

Whole-School Reform and Early Adolescent Development

Andreas de Barros
Alejandra Campos Quintero
Najiba El Amrani Mida
Paul Glewwe
Nikhil Kumar
Laure Lépine

March 25, 2026

Beyond Basics: Whole-School Reform and Early Adolescent Development[*]

Andreas de Barros[†]

Alejandra Campos Quintero[‡]

Najiba El Amrani Mida[§]

Paul Glewwe[¶]

Nikhil Kumar[||]

Laure Lépine[**]

March 25, 2026

This study investigates whether a government-led, whole-school reform in public lower secondary schools can simultaneously reduce dropout and improve both academic and socioemotional outcomes. We employ a prospective difference-in-differences design, using machine learning on nationwide administrative data to match 200 reform schools in Morocco to 100 comparable schools. The analysis uses primary assessment data for 20,036 students and administrative enrollment records for 637,587 observations. After one year, the reform reduced end-of-year dropout by 1.6 percentage points (a 31.4 percent decline), increased learning by 0.52 standard deviations (a 3.3-fold acceleration), and improved a pre-specified index of socioemotional skills by 0.22 standard deviations. These findings demonstrate that government-led interventions can deliver multidimensional benefits during the critical lower secondary school years and highlight the potential of whole-school reform to support adolescent development.

[*] We thank the Ministry of National Education Preschools and Sports in Morocco for supporting the impact evaluation of the Pioneer School Program in lower secondary schools. We are grateful to the Morocco Innovation and Evaluation Lab (MEL)—especially Youssef Assarssah, Arthur Bluche, Ayoub Chaabi, Florencia Devoto, Sara Ameziane Hassani, Zakaria Mansouri, and Andrea Salem—for their excellent field coordination, research assistance, and project leadership. We thank Quentin Daviot and Thomas Escande 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 the University of California, Irvine, Harvard University, and the University of Minnesota. Following common journal policies, we acknowledge the use of Artificial Intelligence (AI) for AI-assisted copy editing, yet did not use AI for generative editorial work and autonomous content creation. We have no conflict of interest to declare. The order of authors is alphabetical; co-authorship is shared equally among the authors.

[†] Assistant Professor, School of Education and (by courtesy) Department of Economics, University of California, Irvine. E-mail: adb@uci.edu.

[‡] PhD Student, Teachers College, Columbia University. E-mail: aac2271@tc.columbia.edu.

[§] PhD Student, Department of Economics, European University Institute. E-mail: najiba.amrani@eui.eu.

[¶] Regents Professor, Department of Applied Economics, University of Minnesota. E-mail: pglewwe@umn.edu.

[||] PhD Student, Harvard University. E-mail: nikhil_kumar1@g.harvard.edu.

[**] Senior Research Associate, Morocco Innovation and Evaluation Lab, University Mohammed VI Polytechnic. E-mail: llepine@povertyactionlab.org.

Keywords: Education; human capital; Morocco; socioemotional skills.

JEL classification: I20, I21, I28, O15.

Study pre-registration: This study and its analyses were registered with a publicly available pre-analysis plan (osf.io/g54kd).

Contents

1 Introduction

Despite significant attention to foundational literacy and numeracy in the early primary grades, a large share of students in low- and middle-income countries (LMICs) enter secondary school without mastering even basic academic skills (de Barros and Ganimian, 2023). The lower secondary school years—early adolescence—are also a period of heightened dropout risk and rapid socioemotional development, both of which have long-term consequences for educational attainment, labor outcomes, and life trajectories (Deming, 2017; Danon et al., 2024).

One promising policy response is whole-school reform: the idea that improving educational outcomes requires a comprehensive redesign of a school's instructional model, culture, and socioemotional support systems—particularly for students who are lagging academically or at risk of dropping out. Yet, rigorous evidence on the effectiveness of such multi-pronged interventions remains limited, especially in LMICs and when implemented by governments at scale. In particular, little is known about how to develop socioemotional skills through public school systems, despite the growing recognition of their importance for long-term well-being and life outcomes.

This study presents new quasi-experimental evidence on the impact of a government-led, large-scale, whole-school reform in the lower secondary school grades. We investigate whether the reform increased student learning, reduced dropout, and enhanced socioemotional outcomes. The "Pioneer School Program" (PSP), launched in 2024 in 232 public lower secondary schools in Morocco, combines multiple evidence-based strategies, targeting students in all three grades of lower secondary school (grades 7-9). Conceptualized as a whole-school reform effort with multiple interlinked components, the reform integrates structured pedagogy, targeted remediation at the beginning of the school year, extracurricular activities, an early-warning system that identifies students at risk of dropping out, tutoring sessions for these at-risk students, and a school-level quality certification system. In a random subset of 84 schools, students in grades 7 and 8 are also provided with socioemotional development workshops. Implemented entirely by the government without support from international organizations or NGOs, the program is now the government's flagship lower secondary school reform, and in 2025 it was scaled up to an additional 535 public lower secondary schools.

As with the evaluation of any intervention, credible estimates of the reform's impacts require a valid counterfactual. To this end, our prospective, pre-registered study uses a difference-in-differences strategy. First, leveraging administrative data on the universe of public lower secondary schools in the country, we employed machine learning methods to identify 100 similar non-participating lower secondary schools (henceforth referred to as comparison schools) that closely resemble 200 (of the 232) lower secondary schools participating in the first year of the reform (henceforth referred to as Pioneer schools). Then, to estimate the program's impact, we collected primary assessment and one-on-one interview data from a random subsample of 20,036 students in these 300 schools and contrasted changes in learning and socioemotional development between the two sets of schools. Additionally, we used administrative data on dropout for all students enrolled in these 300 schools over a two-year period (637,587 student-by-year observations). The key identifying assumption is that the average (conditional) trend in Pioneer schools would have mirrored that of the comparable non-Pioneer schools in the absence of the program. Comparisons of trends in exam scores over the three years preceding the reform support this "parallel trends" assumption.

After one school year, we find that the program led to large average impacts in the desired directions on student dropout, learning outcomes, and socioemotional skills. These effects also hold for students who were identified at baseline as being at risk of dropping out. For this subgroup of students, effects on dropout and socioemotional skills were especially pronounced, and the program also enhanced their creativity.

We present five main sets of results from our pre-registered, quasi-experimental study. First, the program reduced student dropout rates at the end of the school year by 1.6 percentage points. This absolute reduction in dropout reflects a 31.4 percent reduction in dropout (the counterfactual dropout level is 5.1 percent). This effect is driven by a 1.3 percentage point decrease in the percentage of students who drop out voluntarily, vis-à-vis a 0.4 percentage point decrease in the percentage of students who are expelled (or "excluded") due to their continued poor academic performance or their behavioral problems. In addition, the program reduced students' likelihood of repeating a grade by 8.5 percentage points. In comparison with other quasi-experimental studies of educational interventions targeting student dropout in low- and middle-income countries (Evans and Yuan, 2022), these results place the program's effectiveness in reducing dropout within the top 40-50 percent of impacts observed elsewhere.[1]

Second, averaging across Arabic, French, math, and science, the program's impact on student learning is an increase of 0.52 standard deviations (s.d.) of the distribution of test scores of the non-Pioneer schools at the end of the school year. The counterfactual growth in academic skills over the same period was 0.23 s.d., which suggests the program

[1] Focusing on other education programs in Morocco, the reform's effectiveness in reducing dropout rates is comparable to the country's conditional cash transfer program for primary school students ("*Tayssir*"). Benhassine et al. (2015) found that a pilot of Tayssir reduced dropout by 1.3-1.7 percentage points over one year (depending on the treatment arm). These intent-to-treat effects are similar to the local average treatment effects for the national scale-up of the Tayssir program documented by Gazeaud and Ricard (2024), which show a 1.3 percentage point year-on-year reduction.

more than tripled students' rate of learning over the school year (a 3.3-fold acceleration).[2] In comparison with other impacts on student learning observed in low- and middle-income countries (Evans and Yuan, 2022), these results rank the program's effectiveness within the top 20 percent of effects documented by other quasi-experimental studies; when focusing on large-scale (more than 10,000 observations) quasi-experimental studies, it places the program's effectiveness within the top 1 percent of impacts.[3] By subject, the program's effects are increases of 0.24 s.d. in Arabic, 0.31 s.d. in French, 0.30 s.d. in math, and 1.24 s.d. in science. These impacts are consistently positive across fine-grained content and cognitive subdomains, as measured by the assessments we constructed, including written and oral skills, materials at and below students' grade levels, as well as lower- and higher-order thinking skills.

Third, we document positive effects of 0.22 s.d. for a pre-specified, "proximal" outcome index that consists of socioemotional subskills for which the Ministry of Education expected to find larger program effects (consisting of measures of students' growth mindset, self-efficacy, pro-sociality, and emotion perception). Yet, even beyond this proximal outcome measure, we also find positive effects on two broad family indices: interpersonal skills (effect of 0.14 s.d.) and intrapersonal skills (effect of 0.26 s.d.). Except for emotion perception, we find positive impacts on all measures of subdimensions included in these two indices, including pro-sociality (0.20 s.d.), perceived control (0.28 s.d.), self-regulation and discipline (0.16 s.d.), growth mindset (0.20 s.d.), locus of control (0.45 s.d.), self-efficacy (0.11 s.d.), grit (0.18 s.d.), and work discipline and diligence (0.18 s.d.). In contrast, for the overall sample, on average, we do not find statistically significant impacts on a separate measure of student creativity that is not included in these two indices.

Fourth, our pre-registered analyses of subgroup effects focus on our measures of dropout, overall student learning, the proximal measure of socioemotional skills, and creativity. Our pre-analysis plan prescribes the analysis of effects for at-risk students first, followed by subsequent analyses for male students, and then additional exploratory analyses (including female students and other subgroups). Our findings for at-risk students indicate a 3.6 percentage point reduction in dropout (compared to a 15.4 percent counterfactual rate), a 0.45 standard deviation increase in student learning, a 0.34 standard deviation increase

[2] Based on learning gains in the comparison group, we estimate that students' academic skills would have grown by 0.23 standard deviations in the absence of the program. Adding the intent-to-treat effect of 0.52 standard deviations to the observed endline scores implies a total gain of 0.75 standard deviations, corresponding to a 3.26-fold acceleration in student learning relative to the counterfactual.

[3] We may also compare the counterfactual growth of 0.23 s.d. with the yearly learning rate from other countries. For example, in the United States, the achievement norms for lower-secondary grade growth in math are very similar, at 0.34 s.d. in grade 7, 0.28 s.d. in grade 8, and 0.18 s.d. in grade 9, respectively (NWEA, 2025). Across subjects, without the program, our Moroccan sample exhibited substantial heterogeneity in student learning, with the lowest growth in math (0.03 s.d.) and the highest growth in science (0.38 s.d.).

in the study's proximal measure of socioemotional skills, and a 0.28 standard deviation increase in students' creativity. We report detailed results for all the remaining subgroups below, in the main text.

Fifth, the above-mentioned large, positive impacts represent intent-to-treat (ITT) effects, reflecting a school's *assignment* to the intervention. Yet, as with any large-scale, multi-component intervention, it is reasonable to expect the *observed* implementation quality to be imperfect. In fact, 96 percent of students in the program schools participated in remedial instruction at the beginning of the school year, and we document a large, 66 percentage-point increase in the proportion of students taught with explicit instruction. Two-thirds of the students in Pioneer schools participated in extra-curricular activities (a 20 percentage-point increase), and two-thirds of the students who were offered socioemotional workshops participated. However, among the at-risk students targeted for remedial tutoring, we observe only a small 9 percentage point increase in students' participation in tutoring sessions (in the comparison group, 60 percent of at-risk students also received tutoring). Overall, students were 35 percentage points more likely to know their school's social worker, and we document a 0.12 s.d. increase in an index of student-reported school climate and well-being. These numbers—largely positive, albeit imperfect—reflect the challenges of implementing whole-school reforms with fidelity at scale. At the same time, they also suggest the large, positive ITT effects mask even larger treatment-on-the-treated effects.

To assess the robustness of our findings, we conducted a series of pre-specified checks addressing key threats to internal validity and concerns about multiple hypothesis testing. First, since the intervention reduced dropout, it is unsurprising that attrition at follow-up was 6.5 percentage points lower in program schools than in comparison schools. Anticipating this, we pre-registered bounding and reweighting strategies (following Behaghel et al. (2015) and Molina-Millán and Macours (2025)), which confirm that our results are robust to differential attrition. Second, because socioemotional outcomes rely on student self-reports, we took steps to account for potential experimenter demand effects (following Dhar et al., 2022). At the beginning of the study, we measured students' propensity to give socially desirable responses, which allows us to show that the results are not driven by this factor. Third, to mitigate concerns about multiple hypothesis testing, we publicly pre-registered a structured hierarchy of hypotheses and report q-values throughout (following Vivalt et al., 2024). Because of our large sample size and the resulting statistical power, all primary results are precisely estimated. Finally, we also show that the results regarding student dropout are robust to alternative sample definitions.

Our study's first and most direct contribution lies in advancing the understanding of comprehensive educational interventions that integrate multiple program components to advance a "whole-school" reform model. Prior studies often examine these components in isolation, which may overlook the potential benefits of combining these strategies. By investigating a program that includes multiple program components, we are able to provide insights into the overall effectiveness of an integrated approach.[4] Although we do not have a factorial design to separately identify the effects of each component and isolate their complementarities (as in Mbiti et al., 2019), the substantial improvements observed suggest that the combination of these strategies can be effective in enhancing student learning, reducing dropout, and promoting socioemotional skills. This highlights the possible advantages of multi-pronged whole-school reforms that address multiple facets of targeted remediation, structured pedagogy, school climate, and socioemotional support.[5]

The study's second contribution is to the nascent literature on government-led, school-based interventions in low- and lower-middle-income countries that aim to promote educational outcomes that go beyond students' foundational skills, including governments' attempts to foster higher-order thinking, improve socioemotional skills, and reduce dropout.[6] Prior work in this area has studied school-based interventions 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).[7] To the best of our knowledge, this is the first large-scale evaluation of a government-led, socioemotional skill intervention implemented in public schools in a low- or lower-middle-income country.[8]

The study's third contribution is to focus on early adolescence (grades 7 to 9), 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

[4] Ibrahim et al. (2024) provides a recent evaluation of Morocco's "Pioneer School Program" in primary schools, which combines two of the main program components that have been implemented in Morocco's lower secondary schools. That paper does not provide estimates for impacts on socioemotional skills. For reviews of "comprehensive school reform" approaches from the United States, see Desimone (2002) and Borman et al. (2003).

[5] For evidence on the advantages of multi-faceted "big-push" interventions, from outside of education, see Banerjee et al. (2015).

[6] 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).

[7] 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).

[8] In a parallel, ongoing randomized evaluation, Kevin Carney and Avinash Moorthy are currently evaluating the effects of a "happiness curriculum" in public schools in Tripura, India. Brown et al. (2025) study the effect of Indian students' exposure to sustained academic activity on their cognitive endurance. Wang et al. (2016) evaluated a smaller, socioemotional skills program in 35 public schools in rural China. The Chinese program produced mixed results, reducing dropout in lower secondary school after eight months, but with impacts fading out after 15 months.

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; Dam et al., 2025), 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 the formative early adolescence stage.

2 Context and Intervention

2.1 Context

Public provision of education, as governed by the Ministry of Education (*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 percent 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 lower secondary level, in the 2023 TIMSS math assessment, Morocco's eighth graders ranked second to last out of 42 countries, and 64 percent of grade 8 students were below the lowest international benchmark in mathematics (von Davier et al., 2024).

2.2 The Pioneer School Program in Morocco's lower secondary schools

This study evaluates the impact of an innovative education intervention, the "Pioneer School Program," which is being implemented in public lower secondary schools in Morocco. The aim of the program is to improve students' skills in Arabic, French, mathematics, and science. Another aim is to improve students' socioemotional skills, including their self-perception and well-being at school. More broadly, the Ministry of Education expects that positive impacts on students' academic and socioemotional skills will reduce their propensity to drop out of school.

The Pioneer School Program (PSP) was launched in 232 of Morocco's 2544 public lower secondary schools in September 2024.[9] The Ministry considers the program to be its flagship intervention in lower secondary schools, and it scaled the program up to another 535 schools in the 2025-26 school year, covering approximately 32 percent of the students in the country. The remaining 1,777 lower secondary schools are expected to be reached by 2028. The program targets students in all three grades (7, 8, and 9) of lower secondary schools, which corresponds to the seventh, eighth, and ninth years of schooling.[10]

Conceptualized as a whole-school reform effort, the reform comprises multiple interlinked components. The Pioneer School Program in lower secondary schools consists of: (1) a "remediation" period of two months at the beginning of the school year, which is inspired by the "Teaching at the Right Level" ("TaRL") approach; (2) detailed scripted lessons provided to lower secondary school teachers ("structured pedagogy"); (3) training teachers to provide targeted instruction to students; (4) extra-curricular activities after school; (5) implementing a monitoring unit to closely follow students at risk of dropping out; (6) group-administered tutoring sessions for these at-risk students; and (7) a system of "quality certification" of schools.[11] In addition, in a randomly selected subset of 84 of the Pioneer schools, each school's social specialist is expected to give four socioemotional workshops to students in grades seven and eight to support their socioemotional skills.

3 Research Methods

To estimate the overall effect of Morocco's whole-school reform in lower secondary schools, we implemented a prospective difference-in-differences design comparing 200 treatment ("Pioneer") schools with 100 matched comparison schools. Comparison schools were selected using machine learning methods to match treatment schools on pre-intervention administrative characteristics, drawing on nationwide administrative data from before the program launched.

[9] The launch in lower secondary schools comes on the heels of launching the PSP program in public *primary* schools one year prior. A separate study evaluates the effectiveness of this primary school project.

[10] In math and science, the program has started only in grade 7; over the coming years, it will gradually expand to grades 8 and 9.

[11] The program's first two components are inspired by the Global Education Evidence Advisory Panel (GEEAP, 2023). More specifically, the PSP combines two of the panel's "great buy" recommendations on how to address the learning crisis in low- and middle-income countries through targeted remediation and structured pedagogy.

To isolate the effect of one specific program component—the socioemotional support workshops delivered by the social specialists working in the schools—we embedded a randomized controlled trial (RCT) within the treatment group. Among the 200 Pioneer schools, 84 were randomly assigned to receive the workshop component, while the remaining 116 did not receive it. This report focuses on the overall impact of the reform. Results from the within-treatment RCT are given in the Appendix and summarized more briefly; they will be presented in greater detail in a companion paper (conditionally accepted at the *Journal of Development Economics*).

To measure impacts on student learning and socioemotional development, we followed a panel of students who were tested at baseline (September 2024) and endline (June 2025) using standardized assessments administered by trained enumerators. To measure (changes in) dropping out, we use administrative records from the year before the intervention launched and from the first year of the program, covering all enrolled students. We also use administrative data to construct school-level covariates for matching, verify the plausibility of the study's identifying assumptions, and assess the study sample's representativeness.

3.1 Sampling and random assignment

3.1.1 Sampling of schools

This study's sample of schools was constructed in two steps, using administrative data on the universe of government lower secondary schools, which includes the 232 lower secondary PSP schools in Morocco.[12] First, using the "post-double-selection" (PDS) methodology (Belloni et al., 2011, 2012, 2014, 2016), 30 variables were identified that were predictive of either test scores or participation in the program. Second, using these 30 selected variables, Mahalanobis nearest-neighbor matching (without replacement) was used to identify pairs of PSP schools that were very similar to one another (in terms of their Mahalanobis distance). For each of the 100 closest pairs (i.e., 200 PSP schools), a matched (non-PSP) control school was found. This gives triplets of schools with two "close" PSP schools, and one matched non-PSP school.

Table 1 compares, in columns (1) to (3), the 300 lower secondary schools in the study's sample of schools with the 2,244 remaining (public) lower secondary schools in Morocco. It also compares, in columns (4) to (6), the 200 program (PSP) lower secondary schools in

[12] At the stage of sampling schools, we had received data on 2,544 public lower secondary schools (that is, all of the public lower secondary schools included in the country's education information management system). We refer to this set of schools as the "universe" of all public lower secondary schools in Morocco.

the study sample with the 32 other PSP lower secondary schools in Morocco. Compared to the 2,244 non-study schools in Morocco, as seen in columns (1) to (3), the 300 study schools have slightly more teachers (32.0 vs. 29.5), are less likely to be rural (34.0 percent vs. 50.2 percent), have more students (894.4 vs. 770.6), have somewhat younger students (13.9 years vs 14.6 years), serve students with slightly higher average primary school passing exam scores (6.95 vs. 6.84), and their lower secondary school passing exam scores are slightly higher (9.11 vs. 8.94).

Turning to differences between the 200 PSP lower secondary schools in the 300 study schools and the other 32 PSP lower secondary schools in Morocco, as seen in columns (4) to (6), the 200 study PSP schools are somewhat larger (32.4 teachers vs. 27.1 teachers), much less likely to be rural (37.0 percent vs. 65.6 percent), have younger students on average (13.9 years old vs 14.6 years old), and have higher primary school passing exam scores (6.94 vs. 6.79) and higher lower secondary school passing exam scores (9.14 vs. 8.75). There are no other significant differences in school characteristics between the 200 PSP lower secondary schools in the study and the other 32 PSP lower secondary schools in Morocco.

Columns (7) to (10) of Table 1 examine balance in school characteristics between the comparison lower secondary schools and the two types of PSP lower secondary schools (with or without the workshop component), as well as balance between the two types of PSP schools. The findings show that the comparison schools are generally quite similar to both types of PSP schools and that the two types of PSP schools are also quite similar to each other. Joint tests (F-tests) reveal no significant differences across these comparisons.

3.1.2 Sub-sampling of students

The study's unit of analysis is a student. For our analyses of effects on student dropout, our sample includes all 326,819 students enrolled in the 300 schools in the first year of the program's implementation, as well as the 310,768 students enrolled in the previous school year (637,587 observations in total).[13]

For our analyses of effects on student learning and socioemotional skills, we rely on a subsample of students. In each of the study's lower secondary schools, at baseline, 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.

[13] Individual students are included in both cohorts if they were enrolled in one of the 300 study schools in both years. Our analysis focuses on students enrolled at the beginning of the two school years (in September 2023 and in September 2024, respectively). In our additional robustness checks, presented below, we investigate the sensitivity of results to alternative sample definitions.

Students were sampled to take the assessments in one subject only (Arabic, French, mathematics, or 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 18 students per school (6 students from each of the three grades, i.e. grades 7, 8, and 9). For the written math and science assessments, we subsampled up to 18 students per school for each subject (all attending grade 7). 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.[14]

The study's effective subsample for the baseline assessments consists of 20,036 students across the study's 300 schools, of which 4,869 were tested in Arabic, 4,876 were tested in French, 5,141 were tested in mathematics, and 5,150 were tested in science (see Columns (1) and (2) in Panel A of Table 2). Of these students, 92.4 percent were tracked successfully and took the endline assessments.

For this subsample of students, Column (3) of Table 2 presents baseline means for the study's main outcomes and four student-level characteristics, including attrition rates.[15] Column (4) compares these variables between PSP ("treatment") and non-PSP ("comparison") schools. None of the differences in outcomes is statistically significant at the 5 percent level. A joint test for each set of outcomes fails to reject the null of no significant differences (p-values > 0.1). However, students in treatment schools were 6.5 percentage points less likely to attrit by endline. Appendix Table A1 presents similar comparisons among students observed at endline, yielding comparable findings, despite the difference in attrition rates. Overall, students in PSP and non-PSP schools were well-balanced at baseline, and remain comparable after attrition. In pre-registered robustness checks, presented below, we also confirm that differential attrition does not affect the study's main conclusions.

[14] More specifically, all students in each grade and track (APIC or ASCG; "Année Secondaire Collégial Originel Parcours International" or "Année Secondaire Collégial Général") were randomly split into four groups, one for each subject. Within each of the four groups of students sampled for a given 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, and 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.

[15] We introduce these outcome measures in Section 3.2, below. Each score (or index) is standardized by subtracting the mean and dividing by the standard deviation of the distribution of the endline scores of the students in the comparison group.

3.1.3 Random assignment of schools

The study's sample consists of 300 lower secondary schools (200 PSP and 100 matched non-PSP). Out of the 200 PSP schools, we randomly assigned 84 schools to receive the socioemotional support workshops (and the remaining 116 schools not to receive that program component). The intention was to randomly assign one of the two PSP schools in each triplet to receive the socioemotional support treatment so that 100 schools would receive that treatment. However, because of logistical reasons, only 84 PSP schools could receive the socioemotional support treatment, and hence, the triplets were not used to select the 84 schools to receive these workshops. Instead, within the 200 PSP schools, a principal component was calculated based on the average primary school passing exam score, number of teachers, number of students, and fraction of female teachers, which was used to sort the schools into an alternating pattern of triplets and pairs (84 groups in total). Within each group, we then randomly assigned one school to receive the socioemotional support workshops, while the other school (for pairs) or the other two schools (for triplets) did not receive the socioemotional support workshops.

3.2 Measurement

This study is based on: 1. Primary data collected in two data collection rounds; 2. Additional data on implementation quality and take-up; and 3. Administrative data. Primary data collection includes a "baseline" and "endline" in all 300 schools sampled for the study (in September 2024 and June 2025, respectively). During both of these rounds, all students in our study sample completed group-administered, paper-based assessments for one of the four subjects. Those students sampled for Arabic and French also completed one-on-one interviews, capturing their oral language skills and all other non-test measures described below (including those related to socioemotional skills).

Primary data collection was primarily handled by Ministry staff external to the study schools, with the support and oversight of staff employed by the Morocco Innovation and Evaluation Lab (MEL). We adhered to strict data collection protocols, including spot checks and accompaniments, as well as monitoring and debriefs for enumerators (see Glennerster, 2017; J-PAL, 2017).

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.[16] For survey instruments with ordinal response categories, we use the IRT graded response model (GRM). For instruments with strictly

[16] See Jacob and Rothstein (2016) for an accessible introduction to Item Response Theory for economists.

binary items (including responses to test questions, which are classified as correct or 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.[17] For select measures (indicated below), we further aggregate IRT scores into indices by calculating the inverse covariance-weighted average across these scores. Each score (or index) is standardized by subtracting the mean and dividing by the standard deviation of the distribution of the endline scores of the students in the comparison group.

3.2.1 Dropout

Morocco's Ministry of Education granted us access to student-level enrollment data for two years, including information on whether students who were no longer in their initial school were still enrolled because they had switched to other public schools. Using these data, we observed whether any given student had left the country's school system before the end of a given school year. We also obtained information on whether a student left voluntarily, was expelled, or had to repeat a grade. Our main, "global" measure of student dropout is whether a student either left school voluntarily or was expelled[18]. We also complement these data with information on whether students re-enrolled in school in the following (2025-26) school year.

3.2.2 Academic skills

Our study includes four subsamples of students; one for each academic subject (see subsection 3.1.2). In our analyses of effects on overall academic skills, before reporting on subject-wise results, we stacked the four subsamples (for Arabic, French, mathematics, and science) and used their respective standardized test scores as the outcome.

We measure students' mathematical skills across five content domains: numbers, geometry, algebra, data and probability, and measurement. These domains, which are 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

[17] 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.

[18] We include in this category students whose status at the end of the school year could not be determined

science. These domains align with the TIMSS and PISA frameworks, ensuring that the assessment reflects internationally recognized benchmarks for scientific literacy. As with 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.

3.2.3 Socioemotional skills

Interpersonal skills. Our global indicator of *interpersonal* skills consists of a social skills index that combines measures of students' pro-sociality and emotion perception. More specifically, it combines the pro-sociality subscale of the Strengths and Difficulties Questionnaire (SDQ) with 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, 2025).

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, work 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 to delay 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.

Index of proximal socioemotional skills. We consulted with the Ministry of Education to identify (and pre-register) a subset of "proximal" outcome measures of socioemotional skills that the program implementers expected to be most closely aligned with the intervention. We then prespecified the inverse covariance-weighted average of the selected subdomain measures as our main outcome related to students' socioemotional skills (prioritizing it above the two broader summary measures of intrapersonal and interpersonal skills). This "proximal" outcome measure is comprised of the (above-mentioned) indicators of the following socioemotional skills: growth mindset, self-efficacy, pro-sociality, and emotion perception.

3.2.4 Creative thinking skills

We measure fluency, originality, and elaboration in creative tasks using the Torrance Tests of Creative Thinking (TTCT) instrument (Torrance, 1968, 1998), which is 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.

3.2.5 Additional outcomes

School climate and well-being. We construct an index of student school climate and 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.

Out-of-school activities. We collected data on students' activities 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 per day on homework, and an analogous question on time spent doing extracurricular activities after school.

3.2.6 Non-outcome measures

Take-up and implementation quality. During the one-on-one student interviews conducted at endline, we also collected data on whether students (in both treatment and control schools) knew their school's social specialist, whether they had met with the specialist, and whether they had attended a workshop with the specialist. We also asked whether they had participated in a workshop with the school's social specialist since returning to school from the Aïd al-Fitr holiday at the end of Ramadan. This holiday provided a clear marker that students could easily remember, rendering an approximately two-month-long recall period.

Social desirability. To be able to test (and account) for the potential presence of experimenter demand effects, we also measured students' propensity to give socially desirable responses (following Dhar et al., 2022). To this end, we administered a short survey module during one-on-one interviews and constructed the Marlowe-Crowne social desirability scale (Crowne and Marlowe, 1960). This 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 Morocco (for the years 2021, 2022, and 2023). These data provide 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 their academic performance at the time they graduated from primary school).

3.2.7 Psychometric properties

Using data from the study's baseline, Appendix Table B1 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).[19]

In Arabic, French, and science, students attempted almost all test questions, leaving only 1, 4.1, and 3.1 percent of the questions unanswered (see column 5). For math, this percentage is only slightly higher, 5.4. 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).[20] In addition, the tests proved to be internally consistent, with average item-test correlations ranging from 0.28 to 0.38 (see column 7).

The average conditional reliability, reported in column (10) of Appendix Table B1, measures the precision of each instrument across the spectrum of respondent abilities.[21] 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

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

[20] 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.

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

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

where N is the sample size, $\mathrm{Var}(\hat{\theta})$ is the variance of the latent proficiency estimates in the sample, and $\mathrm{SE}(\hat{\theta}_i)$ is the standard error of measurement for student i. This measure reflects the proportion of observed score variance attributable to true differences in ability. 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.

challenging 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 or higher than 0.8. In contrast, the remaining indexes had reliability close to 0.7.

3.3 Analytical strategy

3.3.1 Main approach

This report presents estimates of the intent-to-treat effect of the Pioneer School Program on the outcomes of interest using a strategy that combines difference-in-differences with matching methods, which compares changes over time in the outcome variables of interest for lower secondary schools that were assigned to receive that intervention with the same changes in their matched comparison schools. For all outcomes, the following empirical specification is used:

$$\begin{aligned} Y_{igsjt} = {} & \beta(\mathrm{TREAT}_{sj} \times \mathrm{POST}_t) \\ & + \delta(X_{igsj} \times \mathrm{POST}_t) + \zeta(\eta_j \times \theta_g \times \mathrm{POST}_t) + \gamma_{igsj} + \epsilon_{igsjt} \end{aligned} \tag{1}$$

Here, Y_{igsjt} is outcome Y for student i in grade g in school s in matched triplet j in period t. $TREAT_{sj}$ is the assignment of school s to receive the program, and $POST_t$ is a dummy variable equal to 1 if the outcome Y is measured at endline (i.e., after the intervention was rolled out) and 0 if measured at baseline (i.e., before the intervention was rolled out).

The γ_{igsj} term is a student fixed effect, which measures the impacts on Y_{igsjt} of all student characteristics, observed or unobserved, that do not change over time (in our analyses of outcomes related to student dropout, we replace these student fixed effects with grade-by-matched triplet fixed effects). However, it is possible that the impacts of some student characteristics (which include characteristics of the schools they attend) change over time, so a vector of observed student (and school) characteristics, denoted by X_{igsj}, is added and multiplied by $POST_t$. The corresponding vector of coefficients for these interaction terms, denoted by δ, measures the change in the impacts of these observed student (and school) characteristics over time. Put another way, δ measures the contributions of observed student (and school) characteristics to changes in the outcome variables over time.[22] The student variables available in the dataset are students' age, gender, and primary school passing exam score. The school characteristics are school-level average lower secondary school passing exam score in the previous year, school-level average primary school passing exam scores for students, an indicator for whether a school is in a rural area, its number of students, the number of teachers, the average age of students, and the fraction of female students (all measured at baseline). A LASSO procedure is used to determine in a data-driven way which $X_{igsj} \times POST_t$ interactions to include in the specification.[23]

The three-way interaction $\zeta(\eta_j \times \theta_g \times \text{POST}_t)$ represents grade-by-matched-triplet interactions with the post-period indicator. Intuitively, the inclusion of these fixed effects accounts for how much, on average, a treatment student's matched-triplet comparison peers (who are enrolled in the same grade level) learned between baseline and endline.

The coefficient of interest is β, which captures the intent-to-treat (ITT) effect of assignment to the program. By estimating ITT effects, we allow for the (plausible) possibility that the program was not fully implemented in at least some of the pioneer lower secondary schools (none of the non-pioneer schools received the program). Following de Chaisemartin and Ramirez-Cuellar (2024) on clustering in paired and small-strata experiments, we cluster standard errors at the matched-triplet level.

For β in Equation (1) to estimate the causal impact of the PSP program on the outcome(s) of interest, the identifying assumption is that, conditional on the "control" variables in that equation, the average trend in the outcome(s) over time in the PSP schools would have been the same as the average trend in the matched non-PSP schools if the PSP had not been implemented in the PSP schools. While this assumption cannot be directly tested, one piece of supportive evidence is shown in Figure 1. It shows that average scores from lower secondary school passing exams in these two sets of schools are very similar in the three school years before the program was implemented.

3.3.2 Estimating effects for outcomes without baseline information

Two of the outcomes we included in the endline assessments were not measured at baseline (the measures of prosociality and of chemistry). Hence, we cannot use equation (1). For

[22] Since nearly all students stayed in the same school between baseline and endline data collection, student fixed effects also largely absorb school characteristics.

[23] More specifically, we first calculated each student's change in test scores between the baseline and endline assessments. Then, we residualized this change in test scores, subtracting the comparison group's grade-by-matched-triplet mean. Lastly, with these residuals, we used a post-double-selection LASSO to select the relevant predictors X_{igsj} (Belloni et al., 2011).

these outcomes, we use an ANCOVA specification of the following form:

$$\begin{aligned} Y_{igsj} = \beta \text{TREAT}_{sj} + \delta X_{igsj} + \\ \zeta(\eta_j \times \theta_g) + \epsilon_{igsj} \end{aligned} \tag{2}$$

To the extent that these two outcomes are included in indices constructed using inverse-covariance matrix weighting, we compute the baseline index by preserving the relative weights implied by the covariance matrix observed at endline.

3.3.3 Preregistered subgroup analyses and effects by program variant

To report on program effects for the preregistered student subgroups and to estimate effects by program variant, we re-estimate equations (1) and (2) for the respective sub-samples of students and schools (e.g., for at-risk students only, or by excluding the group of schools with (or without) the workshop component of the program).

3.3.4 Multiple hypothesis testing

Accounting for a pre-registered hierarchy of research hypotheses, we adjust for multiple hypothesis testing by computing the sharpened false discovery rate-adjusted q-values using the Benjamini and Hochberg (1995) correction. Following Vivalt et al. (2024), we place our hypotheses into tiers, which correspond to our prioritization of tests and their sequential implementation.[24] Our analyses of intervention take-up and implementation fidelity are largely based on descriptive statistics, and we do not account for these analyses in our adjustments for multiple hypothesis testing.

4 Results

4.1 Main results

The study's main results relate to four families of outcomes. In order of hierarchy, the respective summary outcome measures for these families are student dropout by the end of the school year, overall test scores (stacked across the four subjects), the proximal measure of socioemotional skills, and creativity. Table 3 and Table 4 present the estimated intent-to-treat (ITT) impacts of the PSP program on each of these outcomes as well as

[24] Our pre-registration (on the Open Science Framework website) includes an Excel file with the complete list of hypothesis tests, along with their sequence and respective adjustments.

on additional subdomains. Given the high precision of our estimates, in the discussion that follows, we mention the FDR-adjusted q-value in the text only when it exceeds 0.05; otherwise, we do not indicate statistical significance.

Panel A of Table 3 shows that the program reduced student dropout rates at the end of the school year by 1.6 percentage points. This absolute reduction in dropout reflects a 31.4 percent reduction in dropout (the counterfactual dropout rate is 5.1 percent). This effect is driven by a 1.3 percentage point decrease in the percentage of students who drop out voluntarily, vis-a-vis a 0.4 percentage point decrease in the percentage of students who are expelled (or "excluded") due to continued poor academic performance or behavioral problems. In addition, the program reduced students' likelihood to repeat a grade by 8.5 percentage points (a 42 percent reduction).

Panel B of Table 3 shows that, averaging across Arabic, French, math, and science, the program's impact on student learning is 0.52 standard deviations (s.d.) of the distribution of test scores from non-Pioneer schools at the end of the school year. The counterfactual growth in academic skills over the same period was 0.23 s.d., which suggests the program more than tripled students' rate of learning over the school year (a 3.3-fold acceleration). By subject, the program's effects are 0.24 s.d. for Arabic, 0.31 s.d. for French, 0.30 s.d. for math, and 1.24 s.d. for science.

For improved comparability, we also link our Moroccan assessment data to international scales from TIMSS, PIRLS, and PISA.[25] Appendix Table A2 presents the intent-to-treat (ITT) effects of the Pioneer School Program after linking to these international frameworks. The results are qualitatively similar to the subject-wise impacts presented above, but can also be interpreted relative to the performance of other countries from the most recent international assessment rounds. In Arabic, a 17.90-point higher score on the most recent PISA language exams would have corresponded to a five-rank higher performance for Morocco. In math, we find a 32.31-point gain relative to the Grade 4 TIMSS scale. In science, the 77.01-point increase is even larger, indicating a substantial improvement.

Panel A of Table 4 documents positive effects of 0.22 s.d. for the pre-specified, "proximal" outcome index that consists of socioemotional subskills for which the Ministry of Education expected to find larger program effects (consisting of measures of students' growth mindset, self-efficacy, pro-sociality, and emotion perception). Yet, even beyond this narrower

[25] We used Item Response Theory (IRT) models to place the national test scores from Morocco on the same latent metric that these international studies use. We included in our assessments anchor items—questions that we adopted directly from publicly released TIMSS, PIRLS or PISA assessment items. Using the item parameters from the TIMSS, PIRLS or PISA, we predicted student ability estimates and transformed them to the international scales, setting the mean to 500 and the standard deviation to 100, consistent with the conventions used by TIMSS/PIRLS and PISA.TIMSS and PIRLS are expressed in 4th Grade scale.

outcome measure, we also find positive effects on two broad indices of interpersonal skills (effect of 0.14 s.d.) and intrapersonal skills (0.26 s.d.), respectively.

For the overall sample, on average, Panel B of Table 4 only shows a small, statistically insignificant impact on the measure of student creativity (effect of 0.05 s.d., $q = 0.34$).

4.2 Subgroup effects

Our analyses of subgroup effects focus on the summary measures of dropout, student learning, the proximal measure of socioemotional skills, and creativity. Our pre-registration prescribes the analysis of effects for at-risk students first, followed by subsequent analyses for male students, and then additional exploratory analyses (including female students and other subgroups).

Panel A of Table 5 shows that the program lowered at-risk students' likelihood of dropping out of school by the end of the year by 3.6 percentage points (compared to a 15.4-percent counterfactual rate) and increased their academic learning by 0.45 standard deviations. Notably, the program improved their socioemotional skills by 0.34 standard deviations (which is higher than the average effect of 0.22 s.d.). Among at-risk students, the reform also led to a 0.28 standard-deviation improvement in students' creativity.

Panel B of Table 5 indicates that there is only limited heterogeneity in program effects by gender. For male students, the program lowered their dropout propensity by 2.6 percentage points (compared to a 7.5-percent counterfactual rate), which is higher than the average effect of 1.6 percentage points. The remaining effects are similar to the average effects for the overall sample. The program increased male students' academic learning by 0.47 standard deviations, and it improved their socioemotional skills by 0.26 standard deviations. As in the overall sample, among male students, we only find a small, statistically insignificant impact on the measure of student creativity (effect of 0.04, $q = 0.453$).

In addition to the above pre-registered investigation of program impacts among at-risk and male students, we also explored additional heterogeneity in effects along a broader set of student and school characteristics. For the analysis of heterogeneity in impacts on dropout, we use administrative data; for the remaining outcomes, to construct a common sample with available data across the three grades, we limit this exploration to the subsample of students selected for one-on-one interviews (who sat the written exams in either Arabic or French). More specifically, to better understand who benefited the most from the program, we used the causal forest approach developed by Athey and Imbens (2016) and Wager and Athey (2018).[26] Following Carlana et al. (2022), we use the out-of-bag predictions to categorize students into the top and bottom half of predicted treatment effects (designated as "Weak" and "Strong" groups). We also present the conditional average treatment effect (CATE) for each subgroup, its standard error, and the respective adjusted q-value.

Focusing on individual background characteristics at the student level (Panels A of Appendix Tables A3 to A6), we find evidence that program effects are larger for students at risk of dropping out, students in the bottom quartile of socioemotional skills, and those in the bottom quartile of academic performance at baseline. As some of the program components focused on grades 7 and 8 only (see above), perhaps unsurprisingly, we also find evidence of slightly larger effects in these two grades (vs. grade 9). While these findings point to a clear pattern of greater effects among weaker students, at the school level, the results are less conclusive (Panels B). For instance, we find slightly higher effects on socioemotional skills and creativity in rural schools and regions of lower economic development; however, this pattern is reversed for effects on student learning.

4.3 Implementation quality and mechanisms

The substantial intent-to-treat (ITT) effects reported above reflect the impact of schools' *assignment* to the intervention. Implementation fidelity was high for most program components, but as with most large-scale, multi-component reforms, it was imperfect. Incomplete delivery or limited student exposure likely means that the observed ITT effects understate the treatment-on-the-treated (ToT) effects.[27]

Panel A of Table 6 summarizes implementation quality and potential mechanisms using administrative data on students' participation in remedial instruction and student reports collected at endline. Column (1) presents treatment group means; column (2) shows estimated treatment effects. For the components targeted to subgroups of students—workshops and tutoring—we restrict the analysis to eligible students (students in the first two grades of lower secondary school for workshops and at-risk students for tutoring). Moreover, because only the randomly selected subset of 84 program schools

[26] To accommodate our difference-in-difference setup, following Gavrilova et al. (2023), we implemented this causal forest analysis as follows. First, we calculated each student's change score by subtracting the baseline test score from the endline test score. Second, using the comparison group, for each subject, we regressed the change scores on matched triplet-by-grade fixed effects and the vector of controls identified in the analyses of main effects (presented above). Third, from this regression, we calculated the residuals. Finally, using the "honest" approach (Athey et al., 2019), we trained a causal forest with these residuals and a pre-specified set of covariates, building 50,000 trees, setting the minimum number of treatment and control observations allowed in a leaf to the default value (5), and accounting for the clustering of students within schools.

[27] Given the multidimensional nature of the reform, we do not estimate ToT effects (e.g., via instrumental variables), as no single indicator sufficiently captures full program exposure.

was assigned to provide workshops, Appendix Table A9 also disaggregates the results by program variant.

In program schools, 96 percent of students participated in remedial instruction (Teaching at the Right Level, or TaRL) at the beginning of the school year. In addition, we observe a 66 percentage-point increase in the use of explicit instruction, with 89 percent of students reporting that their teacher had used a projector and slides in the past month. Student-reported participation in extracurricular activities rose by 20 percentage points, reaching two-thirds of students in Pioneer schools. In the subset of schools assigned to deliver socioemotional workshops, 66 percent of eligible students reported attending a group session with the school's social specialist—a 55 percentage-point increase (see Appendix Table A9).[28]

Conversely, the tutoring component for at-risk students was implemented more modestly. Among at-risk students targeted for tutoring, the program increased participation by just 9 percentage points. This result may reflect the already high tutoring rates in the comparison group (60 percent).

As for potential mechanisms, students in program schools were 35 percentage points more likely to report knowing their school's social worker (the "social specialist"). We also find a 0.12 standard deviation improvement in the index capturing school climate and student well-being. By contrast, we detect no meaningful changes in students' out-of-school study behavior (time spent on homework), suggesting that the observed program impacts may have stemmed primarily from improvements in *within*-school productivity.

4.4 Impacts on additional subdomains of academic and socioemotional skills

The program's impacts on student learning are robust and consistently positive across fine-grained content and cognitive subdomains, as measured by the assessments we constructed, including written and oral skills, materials at and below students' grade level, as well as lower- and higher-order thinking skills. Yet, Appendix Table A7 points to some heterogeneity across these subdomains. In Arabic and French (Panels A and B), we find the effects are larger for oral skills compared to students' written production of text (0.31 and 0.35 s.d. vs. 0.12 and 0.19 s.d.). In mathematics (Panel C), we find larger effects for the number sense and operations subdomain (0.46 s.d. vs. the overall effect of 0.30 s.d.). In science (Panel D), effects are particularly large for items requiring lower-order thinking skills ("knowing"; 1.68 s.d.) and in chemistry (2.31 s.d.).

[28] At endline, we asked students about whether they had attended a group session with the social specialist. In control-group schools, 11 percent of students reported having met with their school's social specialist in a group setting.

For socioemotional skills, except for the emotion perception subdomain, we find positive impacts on all measures of subdomains included in the two indices of interpersonal and intrapersonal skills. Appendix Table A8 documents positive effects of the program on students' pro-sociality (0.20 s.d.), perceived control (0.28 s.d.), self-regulation and discipline (0.16 s.d.), growth mindset (0.20 s.d.), locus of control (0.45 s.d.), self-efficacy (0.11 s.d.), grit (0.18 s.d.), and work discipline and diligence (0.18 s.d.).[29]

4.5 Results by program variant

The random assignment of a subset of 84 schools to the socioemotional support intervention in grades 7 and 8 enables us to tease out whether these workshops contribute to the overall positive program effects. Overall, we do not find strong support for their added effectiveness.

Appendix Tables A10 and A11 show the program effects for the two grades by program variant (without or with workshops). These results suggest that the workshops did not further decrease the dropout rate (a 1.6 percentage-point decrease with workshops, compared to a 2.0 percentage-point decrease without workshops). However, the workshops may have slightly reduced students' likelihood of repeating a grade (a 10.2 percentage point decrease, vs. a 9.3 percentage point decrease without workshops). The effects on student learning are similar across the two program variants, with only slightly larger effects for the group with workshops (0.62 s.d. vs. 0.55 s.d.). We do not find benefits from adding the workshops on students' socioemotional skills or creativity; if anything, the program effects on intrapersonal skills and creativity are larger for the group without the workshops.

In exploring possible explanations for this apparent lack of a difference in program impacts, we point to Appendix Table A9, which indicates that exposure to the workshops was imperfect, that students in schools assigned to the workshops were only slightly more likely to know their social specialist, and that schools assigned to the workshops did not observe additional improvements to their school climate or student well-being.

4.6 Robustness checks

To assess the robustness of our findings, we conducted a series of pre-specified checks addressing key threats to internal validity.

[29] Students were randomly subsampled to sit their written exam measuring academic language skills in either Arabic or French. Perhaps unsurprisingly, Appendix Table A8 also shows that the program impacts on socioemotional skills are very similar across these two subgroups of students.

First, since the intervention reduced dropout, it is unsurprising that attrition at follow-up was 6.5 percentage points lower in program schools than in comparison schools. Anticipating this, we pre-registered bounding and reweighting strategies, which confirm that our main results are robust to differential attrition. In the bounding exercise, in turn, we remove either the top or bottom performing students from the group of reform-school students such that the attrition rates balance. As simple Lee (2009) bounds are very imprecise, we follow Behaghel et al. (2015) and trim students according to the time required to successfully track them during the follow-up data collection period. In addition, we calculate inverse-probability-weighted (IPW) estimates of program impacts. As with the bounding exercise, we also present results from a traditional IPW approach and a more modern approach that improves over simple IPW estimates by integrating information on the tracking intensity needed to follow up on students (following Molina-Millán and Macours, 2025). Presented in Panels B and D of Appendix Table A12, the results from these two modern approaches show that both the bounded estimates and the IPW estimates closely match the main results presented in Tables 3 and 4.

Second, because socioemotional outcomes rely on student self-reports, we took steps to account for potential surveyor demand effects (following Dhar et al., 2022). At the beginning of the study, we measured students' propensity to give socially desirable responses, which allows us to confirm that social desirability bias does not drive our results. Table A13 explores heterogeneity in treatment effects on socioemotional skills (and creativity) among the top and bottom halves of students who, at baseline, provided more (less) socially desirable answers. The intuition is that, if experimenter demand effects drove the estimated program impacts, those students with a higher propensity to provide socially desirable answers should exhibit larger effect estimates. While there is some heterogeneity in the treatment effects across the various subdomains, all reported effects remain positive and statistically significant (including for the group of students with lower social desirability scores). Overall, across students with high and low social desirability scores, there is no notable difference in the estimated effects on the main, prespecified outcome measure of socioemotional skills.

Finally, in addition to these pre-specified checks, we also demonstrate that our results regarding student dropout rates are robust to alternative sample definitions. As student enrollment continues at the start of a given school year, with students transferring in and out of schools, it is not immediately clear what population should be tracked from the beginning of the school year. In Appendix Table A14, we show that the program did not lead to systematic transfers of students into program schools (if anything, students were slightly more likely to transfer out of the program schools). Moreover, Appendix Table A15 confirms that the effects on student dropout are very similar to those reported in Table 3, even if we construct the sample of students to be tracked with alternative cutoff dates at the beginning of the school year.

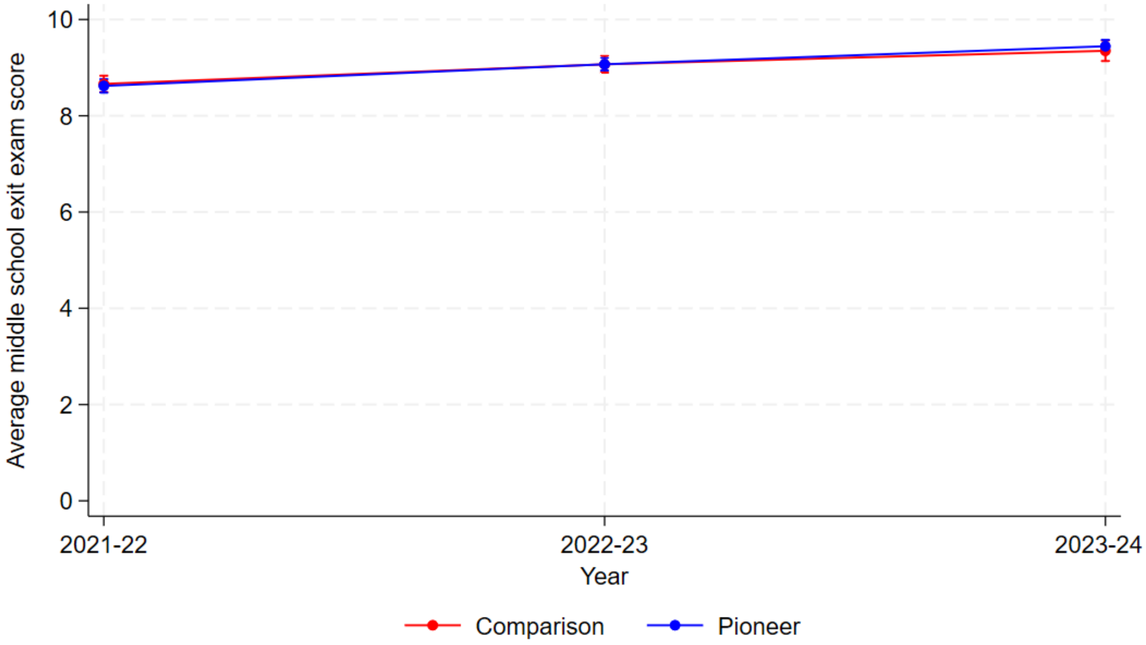
5 Conclusion

This paper provided large-scale quasi-experimental evidence on the impacts of a government-led, whole-school reform in lower secondary schools in a lower-middle-income country. After one year, Morocco's flagship "Pioneer School Program" produced large and precisely estimated improvements in student outcomes: it reduced dropout by nearly one-third, more than tripled the rate of student learning, and enhanced a broad range of socioemotional skills. Effects were especially pronounced for students at risk of dropping out, for whom the program also boosted creativity.

These impacts demonstrate that multi-component "whole-school" reforms, implemented by governments at scale, can reduce dropout and deliver multidimensional improvements in academic and socioemotional outcomes. Our findings expand the small body of evidence on socioemotional skill interventions delivered through public education systems in low- and lower-middle-income countries, demonstrating that such skills can be improved during early adolescence through existing school structures. The results also highlight the potential of lower secondary school as a critical window for interventions aimed at reducing dropout, accelerating learning, and strengthening socioemotional skills.

From a policy perspective, our findings suggest that comprehensive, government-led school reforms can be a promising investment for improving multiple dimensions of student success. Future work could explore the persistence of impacts over time, including potential effects on later-life outcomes, and assess whether similar approaches can be adapted successfully in other contexts.

Figure 1: *Parallel trends*



Notes: This figure confirms that the program and comparison schools exhibited similar levels and similar trends in exam scores before the program started. It presents, for the 200 program schools and 100 comparison schools, the yearly average of lower secondary school passing exam scores over a period of three years before the program launched.

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

	Representativeness (overall)			Representativeness (within program)			Balancing check			
	Non-Study	Study	Difference	Other PSP	Study PSP	Difference	Comparison School mean	Without Socioemotional vs comparison	With Socioemotional vs comparison	With Socioemotional vs without
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Number of teachers	29.53	32.02	2.49***	27.12	32.38	5.25**	31.29	0.55	2.02*	-1.19
	[15.34]	[13.44]	(0.84)	[12.53]	[14.35]	(2.41)	[11.41]	(1.14)	(1.15)	(2.07)
Rural (%)	50.22	34.00	-16.22***	65.62	37.00	-28.62***	28.00	9.64*	6.02	1.89
	[50.01]	[47.45]	(2.93)	[48.26]	[48.40]	(9.10)	[45.13]	(5.18)	(6.03)	(6.96)
Total Enrolment	770.62	894.44	123.82***	834.62	913.21	78.58	856.91	30.53	98.88**	-66.38
	[460.48]	[436.04]	(26.96)	[457.39]	[466.13]	(86.48)	[367.90]	(43.31)	(41.78)	(65.59)
Female students (%)	47.32	47.05	-0.27	48.04	47.09	-0.95	46.97	0.38	-0.17	-0.88**
	[4.98]	[2.89]	(0.20)	[3.53]	[2.90]	(0.65)	[2.89]	(0.44)	(0.53)	(0.41)
Age	14.63	13.94	-0.70***	14.58	13.94	-0.64***	13.93	0.01	0.00	-0.05
	[0.34]	[0.21]	(0.01)	[0.26]	[0.22]	(0.05)	[0.21]	(0.03)	(0.03)	(0.03)
Primary school passing exam score	6.84	6.95	0.11***	6.79	6.94	0.15***	6.97	-0.02	-0.03	0.00
	[0.29]	[0.27]	(0.02)	[0.23]	[0.28]	(0.04)	[0.26]	(0.03)	(0.04)	(0.04)
Lower secondary school passing exam score	8.94	9.11	0.16*	8.75	9.14	0.39*	9.05	0.03	0.09	0.15
	[1.46]	[1.37]	(0.09)	[1.09]	[1.36]	(0.21)	[1.40]	(0.12)	(0.14)	(0.20)
Number of schools	2244	300	2544	32	200	232	100	216	184	200
Joint F-test (p-value)			0.00			0.00		0.63	0.26	0.13

Notes. This table reports on the study's sample of schools. “‘Study” refers to the 300 schools included in the study's effective sample. “Non-study” refers to all other government lower secondary schools in Morocco. “Study (PSP)” and “Other (PSP)” refer to 200 Pioneer schools in the study vs. the remaining 32 Pioneer lower secondary schools in the country. Comparison refers to the 100 matched comparison schools (vs. the 200 Pioneer schools). “With Socioemotional” refers to the 84 schools assigned to receive the workshop component of the program. Difference reports on the regression-adjusted difference (controlling for matched-triplet fixed effects in columns 8 and 9). Standard errors are clustered at the school level (and at the matched-triplet level in columns 8 and 9). Standard deviations are shown in brackets; standard errors are shown in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table 2: *Balance checks*

	Number of observations			Balancing check
	Comparison	Treatment	Comparison mean	Treatment vs. comparison
	(1)	(2)	(3)	(4)
Panel A: Academic skills				
Learning (stacked test scores)	6743	13293	-0.25	-0.03
			[1.18]	(0.04)
Arabic	1646	3223	-0.29	-0.05
			[1.01]	(0.09)
French	1652	3224	-0.23	0.06
			[0.95]	(0.09)
Math	1713	3428	-0.06	0.02
			[1.17]	(0.12)
Science	1732	3418	-0.41	-0.09
			[1.48]	(0.17)
Physics and Chemistry	1732	3418	-0.03	-0.28
			[1.47]	(0.18)
Life Science	1732	3418	-0.81	0.08
			[1.51]	(0.17)
Joint F-test (p-value)				0.11
Panel B: Socioemotional skills and creativity				
Proximal outcome measure	3298	6447	2.03	-0.02
			[0.99]	(0.06)
Interpersonal skills	3298	6447	-0.34	0.01
			[1.29]	(0.09)
Emotional perception	3298	6447	-0.26	0.01
			[0.98]	(0.07)
Intrapersonal Skills	3298	6447	2.57	-0.06
			[0.98]	(0.06)
Personal control	3298	6447	2.22	-0.08
			[1.14]	(0.08)
Self-regulation and discipline	3298	6447	2.25	-0.01
			[0.95]	(0.05)
Creativity	3298	6447	-0.24	0.07
			[1.05]	(0.09)
Joint F-test (p-value)				0.81
Panel C: Student characteristics				
Age	6743	13293	12.88	-0.02
			[1.20]	(0.03)
Primary school passing exam score	6743	13293	7.14	0.00
			[1.17]	(0.00)
% Female	6743	13293	50.87	-2.15
			[50.00]	(1.63)
% Attrited	6743	13293	11.73	-6.47***
			[32.18]	(1.33)
Joint F-test (p-value)				0.00

Notes. This table describes the study's sample of 20,036 students and presents balance checks. Comparison refers to 100 matched comparison schools. "Treatment" refers to the 200 Pioneer schools. "Balancing check" reports on the regression-adjusted difference between Pioneer schools and comparison schools, controlling for matched-triplet-by-grade fixed effects, in column 4. Reversed outcomes were flipped so higher scores represent desirable outcomes. Standard errors are clustered at the matched-triplet level in column 4. Standard deviations are shown in brackets; standard errors are shown in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table 3: *Intent-to-treat effects on student dropout and learning*

	Counterfactual	Main results
	Levels (A) or Growth (B)	Overall ITT effect
	(1)	(2)
Panel A: Dropout and repetition		
Dropout by end of school year (global: either excluded or dropout)	0.051	-0.016***
	[0.003]	(0.003)
		{0.000}
Dropout by end of the school year	0.038	-0.013***
	[0.002]	(0.002)
		{0.000}
Excluded by end of school year	0.013	-0.004*
	[0.002]	(0.002)
		{0.068}
Repeated at end of school year	0.201	-0.085***
	[0.007]	(0.006)
		{0.000}
Long term dropout	0.055	-0.014***
	[0.003]	(0.002)
		{0.000}
Panel B: Learning		
Overall (stacked test scores)	0.23***	0.52***
	(0.03)	(0.03)
		{0.000}
Arabic	0.26***	0.24***
	(0.04)	(0.04)
		{0.000}
French	0.21***	0.31***
	(0.04)	(0.04)
		{0.000}
Math	0.03	0.30***
	(0.04)	(0.04)
		{0.000}
Science	0.38***	1.24***
	(0.06)	(0.06)
		{0.000}
Physics & Chemistry	0.00	1.24***
	(0.06)	(0.06)
		{0.000}
Life Science	0.79***	1.15***
	(0.07)	(0.07)
		{0.000}

Notes. This table reports on the program's intent-to-treat (ITT) effects following equation 1 . Panel A reports the effect of the program on dropout among all students present in the study schools relative to the dropout in the prior year. Our sample includes all 326,819 students enrolled in the 300 schools in the first year of the programs implementation, as well as the 310,768 students enrolled in the previous school year (637,587 observations in total). Column 1 reports the estimated counterfactual level of dropout in comparison schools and column 2 reports the ITT effect. Panel B shows the effect of the program on student learning among the 18,512 non-attriting students across our 300 government lower secondary schools. Column 1 reports the estimated counterfactual growth in the treatment group, and column 2 reports the ITT effect. Standard errors are clustered at the matched-triplet level. Standard deviations are shown in brackets; standard errors are shown in parentheses. q-values are shown in curly brackets, using the Benjamini and Hochberg (1995) correction for multiple hypothesis testing (MHT), and following Vivalt et al. (2024). Main family measures are highlighted in bold font. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, before adjustment for MHT.

Table 4: *Intent-to-treat effects on socioemotional skills and creativity*

	Treatment group	Main results
	Counterfactual growth	Overall ITT effect
	(1)	(2)
Panel A: Socioemotional skills		
Proximal outcome measure	-2.05***	0.22***
	(0.03)	(0.03)
		{0.000}
Interpersonal skills	0.33***	0.14***
	(0.04)	(0.04)
		{0.001}
Pro-sociality		0.20***
		(0.03)
		{0.000}
Emotion perception	0.25***	-0.03
	(0.03)	(0.03)
		{0.315}
Intrapersonal skills	-2.59***	0.26***
	(0.04)	(0.04)
		{0.000}
Perceived control	-2.24***	0.28***
	(0.04)	(0.04)
		{0.000}
Self-regulation and discipline	-2.26***	0.16***
	(0.04)	(0.04)
		{0.000}
Panel B: Creativity		
Creativity	0.22***	0.05
	(0.05)	(0.05)
		{0.340}

Notes. This table reports on the program's intent-to-treat (ITT) effects following equation 1 on students' socioemotional skills and creativity among the 8,959 non-attriting students across our 300 government lower secondary schools who sat the written exams in Arabic or French. Column 1 reports the estimated counterfactual growth in the treatment group, and column 2 reports the ITT effect. Standard errors are clustered at the matched-triplet level. "Prosociality" was not assessed at baseline, hence it does not have a counterfactual growth and its ITT effect is estimated using the ANCOVA specification in equation 2. Standard deviations are shown in brackets; standard errors are shown in parentheses. We show *q*-values in curly brackets, using the Benjamini and Hochberg (1995) correction for multiple hypothesis testing (MHT), and following Vivalt et al. (2024). Main family measures are highlighted in bold font. * p < 0.10, ** p < 0.05, *** p < 0.01, before adjustment for MHT.

Table 5: *Pre-registered analysis of subgroup effects*

	Treatment group	Main results
	Counterfactual levels / growth	Overall ITT effect
	(1)	(2)
Panel A: At-risk students		
Dropout by end of school year (global: either excluded or dropout)	0.154***	-0.036***
	(0.009)	(0.009)
		{0.000}
Learning (stacked test scores)	0.31***	0.45***
	(0.05)	(0.05)
		{0.000}
Socioemotional skills (proximal outcome measure)	-2.22***	0.34***
	(0.09)	(0.09)
		{0.000}
Creativity	0.07	0.28***
	(0.09)	(0.09)
		{0.004}
Panel B: Male students		
Dropout by end of school year (global: either excluded or dropout)	0.075***	-0.026***
	(0.004)	(0.004)
		{0.000}
Learning (stacked test scores)	0.18***	0.47***
	(0.03)	(0.03)
		{0.000}
Socioemotional skills (proximal outcome measure)	-2.20***	0.26***
	(0.05)	(0.05)
		{0.000}
Creativity	0.18***	0.04
	(0.06)	(0.06)
		{0.453}

Notes. This table reports on the program's intent-to-treat (ITT) effects among pre-registered subgroups following equation 1 for the four headline outcomes from tables 3 and 4. We restrict the analysis to subgroups of the same sample of students from the analysis in tables 3 and 4. Panel A reports the effects for at-risk students, and Panel B reports the effects for male students. Column 1 reports the estimated counterfactual growth in the treatment group, and column 2 reports the ITT effect. Standard errors are clustered at the matched-triplet level. Standard deviations are shown in brackets; standard errors are shown in parentheses. q-values are shown in brackets, using the Benjamini and Hochberg (1995) correction for multiple hypothesis testing (MHT), and following Vivalt et al. (2024). Main family measures are highlighted in bold font. * p < 0.10, ** p < 0.05, *** p < 0.01, before adjustment for MHT.

Table 6: *Program implementation, student exposure, and mechanisms*

	Treatment group	Main results
	Overall mean	Overall ITT effect
	(1)	(2)
Panel A: Program implementation and exposure		
Student received TaRL	0.96	0.96***
	[0.20]	(0.01)
Student received explicit teaching	0.89	0.66***
	[0.31]	(0.02)
Students participation in extra-curricular activities	0.66	0.20***
	[0.48]	(0.02)
Student participated in socioemotional workshops	0.34	0.23***
	[0.48]	(0.02)
Student participated in tutoring program	0.53	-0.03*
	[0.50]	(0.02)
At-risk student participated in tutoring program	0.69	0.09**
	[0.46]	(0.04)
Targeted student participated in socioemotional workshops	0.40	0.29***
	[0.49]	(0.02)
Panel B: Mechanisms		
School climate and well-being	0.12	0.12***
	[1.06]	(0.04)
Knows social specialist	0.50	0.35***
	[0.50]	(0.03)
Study habits (spent more than 30 min/day doing homework after school)	0.96	0.01
	[0.19]	(0.01)

Notes. This table describes the program's implementation and exposure to students in the 300 study schools among our sample of 18,512 non-attriting students. "Treatment group" refers to the 200 Pioneer schools. "Main results" reports on the regression-adjusted difference between pioneer schools and comparison schools, controlling for matched-triplet-by-grade fixed effects, in column 2. Reversed outcomes were flipped so higher scores represent desirable outcomes. Standard errors are clustered at the matched-triplet level in column 2. Standard deviations are shown in brackets; standard errors are shown in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

References

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.

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.

Athey, S., Tibshirani, J., Wager, S., 2019. Generalized random forests. The Annals of Statistics 47, 1148–1178. doi:10.1214/18-AOS1709. publisher: Institute of Mathematical Statistics.

Banerjee, A., Duflo, E., Goldberg, N., Karlan, D., Osei, R., Parienté, W., Shapiro, J., Thuysbaert, B., Udry, C., 2015. A multifaceted program causes lasting progress for the very poor: Evidence from six countries. Science 348, 1260799. doi:10.1126/science.1260799. publisher: American Association for the Advancement of Science.

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., Ganimian, A.J., 2023. The Foundational Math Skills of Indian Children. Economics of Education Review 92, 102336. doi:10.1016/j.econedurev.2022.102336.

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.

Behaghel, L., Crépon, B., Gurgand, M., Le Barbanchon, T., 2015. Please call again: Correcting nonresponse bias in treatment effect models. Review of Economics and Statistics 97, 1070–1080.

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.

Benhassine, N., Devoto, F., Duflo, E., Dupas, P., Pouliquen, V., 2015. Turning a Shove into a Nudge? A Labeled Cash Transfer for Education. American Economic Journal: Economic Policy 7, 86–125. doi:10.1257/pol.20130225.

Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) 57, 289–300.

Borman, G.D., Hewes, G.M., Overman, L.T., Brown, S., 2003. Comprehensive School Reform and Achievement: A Meta-Analysis. Review of Educational Research 73, 125–230. doi:10.3102/00346543073002125. publisher: American Educational Research Association.

Brown, C., Kaur, S., Kingdon, G., Schofield, H., 2025. Cognitive Endurance as Human Capital. The Quarterly Journal of Economics 140, 943–1002. doi:10.1093/qje/qjae043.

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

de Chaisemartin, C., Ramirez-Cuellar, J., 2024. At What Level Should One Cluster Standard Errors in Paired and Small-Strata Experiments? American Economic Journal: Applied Economics 16, 193–212. doi:10.1257/app.20210252.

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

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 <span style="font-variant:small-caps;">metaanalysis</span> of universal

<span style="font-variant:small-caps;">schoolbased SEL</span> 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. place: US Publisher: American Psychological Association.

Dam, A., Gray-Lobe, G., Kremer, M., de Laat, J., Morsink, K., 2025. Learning to Work Towards Goals: A Sequential Evaluation of the Effect of Goal-Setting Course on Academic and Soft Skills. Working Paper 25-1344. Annenberg Institute at Brown University. Providence, RI. URL: https://edworkingpapers.com/sites/default/files/ai25-1344.pdf.

Danon, A., Das, J., de Barros, A., Filmer, D., 2024. Cognitive and Socioemotional Skills in Low-Income Countries: Measurement and Associations with Schooling and Earnings. Journal of Development Economics 168, 103132. doi:10.1016/j.jdeveco.2023.103132.

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. 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.

Desimone, L., 2002. How Can Comprehensive School Reform Models Be Successfully Implemented? Review of Educational Research 72, 433–479. doi:10.3102/00346543072003433. publisher: American Educational Research Association.

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. 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.

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.

Gavrilova, E., Langørgen, A., Zoutman, F.T., 2023. Dynamic Causal Forests, with an Application to Payroll Tax Incidence in Norway. Working Paper 10532. CESifo. doi:10.2139/ssrn.4500857.

Gazeaud, J., Ricard, C., 2024. Learning effects of conditional cash transfers: The role of class size and composition. Journal of Development Economics 166, 103194. doi:10.1016/j.jdeveco.2023.103194.

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.

Ibrahim, H., de Barros, A., Deschênes, S., Glewwe, P., 2024. The Best Buy? Prospective Evidence on Successful Remediation in Moroccos Public Primary Schools. Unpublished manuscript. Morocco Innovation and Evaluation Lab. Rabat, Morocco.

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

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.

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.

Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., Rajani, R., 2019. Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania. The Quarterly Journal of Economics 134, 1627–1673. doi:10.1093/qje/qjz010.

Molina-Millán, T., Macours, K., 2025. Attrition in Randomized Controlled Trials: Using Tracking Information to Correct Bias. Economic Development and Cultural Change 73, 811–834. doi:10.1086/730612.

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. doi:10.6017/lse.tpisc.tr2103.kb5342.

NWEA, 2025. 2025 MAP Growth Norms Technical Manual. White Paper 2025.1.0. Northwest Evaluation Association. Portland, OR. URL: https://www.nwea.org/resource-center/white-paper/88182/MAP-Growth-Norms_NWEA_Technical-Manual.pdf/.

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.

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. doi:10.3386/w32719.

Wager, S., Athey, S., 2018. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. Journal of the American Statistical Association 113, 1228–1242. doi:10.1080/01621459.2017.1319839.

Wang, H., Chu, J., Loyalka, P., Xin, T., Shi, Y., Qu, Q., Yang, C., 2016. Can Social-Emotional Learning Reduce School Dropout in Developing Countries? Journal of Policy Analysis and Management 35, 818–847. doi:10.1002/pam.21915.

Weidmann, B., Xu, Y., 2025. Measuring Emotion Perception Ability Using AI-Generated Stimuli: Development and Validation of the PAGE Test. Journal of Intelligence 13. doi:10.3390/jintelligence13090116.

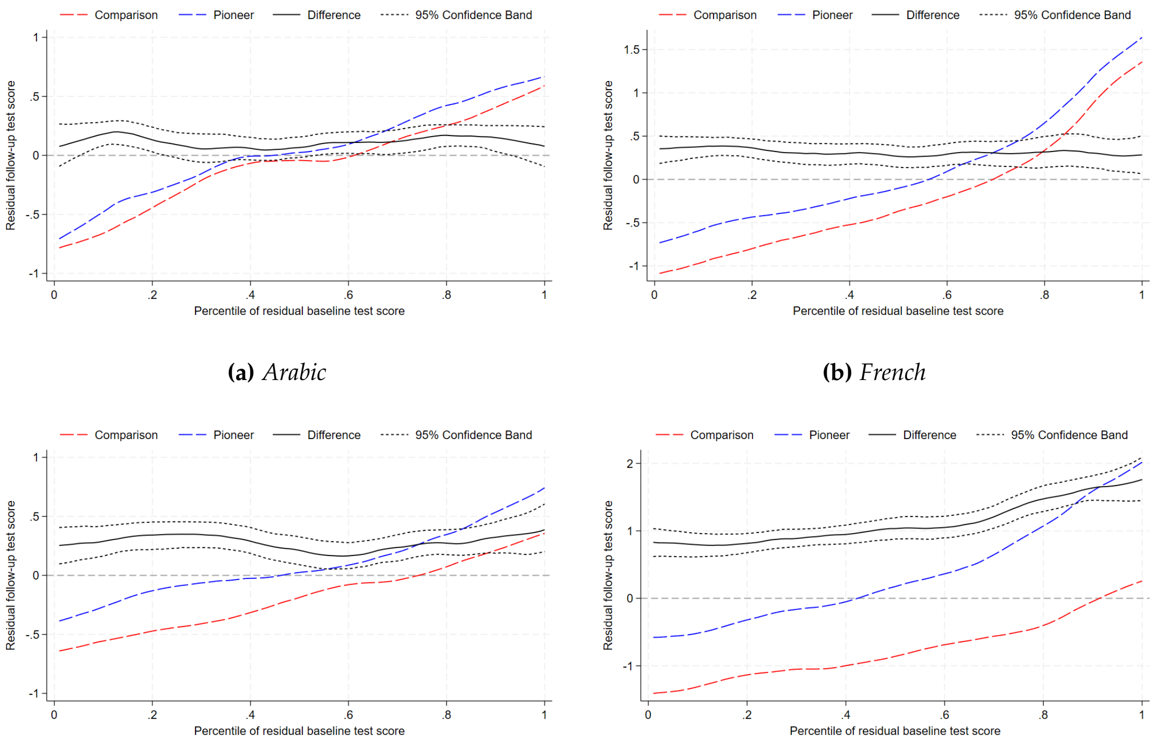
Wilson, S.J., Freeman, B., Hedberg, E.C., 2025. Empirical Benchmarks for Effect Size Interpretation and Study Planning with Social and Behavioral Outcomes. Journal of Research on Educational Effectiveness , 1–26doi:10.1080/19345747.2024.2427767.

Yadlowsky, S., Fleming, S., Shah, N., Brunskill, E., Wager, S., 2025. Evaluating Treatment Prioritization Rules via Rank-Weighted Average Treatment Effects. Journal of the American Statistical Association 120, 38–51. doi:10.1080/01621459.2024.2393466.

Appendix

A Additional Figures and Tables

Figure A1: *Non-parametric investigation of ITT effects by percentile of baseline scores*



(a) *Arabic*

(b) *French*

(c) *Math*

(d) *Science*

Notes: These figures provide a non-parametric investigation of ITT effects by percentiles of baseline scores. The pioneer and comparison lines are estimated using local linear regressions. The pointwise treatment effects are calculated as the difference. The 95% confidence intervals are estimated using bootstrapping; bootstrap iterations are blocked at the matched triplet level, to allow for the clustering of standard errors. The x-axis is the percentile of a students test score at baseline. The y-axis is the residual of a regression of a students test score at follow-up on matched triplet-by-grade fixed effects and a vector of student- and school-level covariates, selected via LASSO. “Baseline” refers to the assessment conducted in September 2024. “Follow-up” refers to the assessment conducted in June 2025

Table A1: *Balance checks (non-attrited sample)*

	Number of observations			Balancing check
	Comparison	Treatment	Comparison mean	Treatment vs. comparison
	(1)	(2)	(3)	(4)
Panel A: Academic skills				
Learning (stacked test scores)	5952	12560	-0.23	-0.02
			[1.18]	(0.05)
Arabic	1460	3032	-0.26	0.00
			[1.01]	(0.10)
French	1453	3014	-0.22	0.03
			[0.95]	(0.11)
Math	1523	3266	-0.04	0.03
			[1.18]	(0.14)
Science	1516	3248	-0.40	-0.04
			[1.46]	(0.17)
Physics and Chemistry	1516	3248	-0.02	-0.25
			[1.46]	(0.19)
Life Science	1516	3248	-0.80	0.13
			[1.50]	(0.17)
Joint F-test (p-value)				0.09
Panel B: Socioemotional skills and creativity				
Proximal outcome measure	2913	6046	2.05	-0.03
			[0.98]	(0.07)
Interpersonal skills	2913	6046	-0.33	-0.01
			[1.29]	(0.10)
Emotional perception	2913	6046	-0.25	-0.01
			[0.97]	(0.07)
Intrapersonal Skills	2913	6046	2.59	-0.04
			[0.98]	(0.06)
Personal control	2913	6046	2.24	-0.07
			[1.14]	(0.09)
Self-regulation and discipline	2913	6046	2.26	0.01
			[0.96]	(0.06)
Creativity	2913	6046	-0.23	0.04
			[1.05]	(0.10)
Joint F-test (p-value)				0.96
Panel C: Student characteristics				
Age	5952	12560	12.86	-0.02
			[1.19]	(0.03)
Primary school passing exam score	5952	12560	7.15	0.00
			[1.12]	(0.00)
% Female	5952	12560	51.18	-2.36
			[49.99]	(1.64)
Joint F-test (p-value)				0.31

Notes. This table describes the study's sample of 18,512 non-attrited students and presents balance checks. Comparison refers to 100 matched comparison schools. "Treatment" refers to the 200 Pioneer schools. "Balancing check" reports on the regression-adjusted difference between Pioneer schools and comparison schools, controlling for matched-triplet-by-grade fixed effects, in column 4. Reversed outcomes were flipped so higher scores represent desirable outcomes. Standard errors are clustered at the school level in columns 5, 6, and 7. Standard deviations are shown in brackets; standard errors are shown in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A2: *Link to international assessments*

	Comparison-group	Main results
	Counterfactual Growth	Overall ITT effect
	(1)	(2)
Panel A: TIMSS / PIRLS		
Arabic	15.61**	33.87***
	(5.86)	(5.86)
		{0.000}
Math	-3.83	32.31***
	(4.54)	(4.54)
		{0.000}
Science	17.43***	77.01***
	(4.35)	(4.35)
		{0.000}
Panel B: PISA		
Arabic	9.03**	17.90***
	(3.89)	(3.89)
		{0.000}

Notes. This table reports on the program's intent-to-treat (ITT) effects following equation 1 . The table shows the effect of the program on student learning among the 18,512 non-attriting students across our 300 government lower secondary schools. Panel A reports the effect on learning outcomes on the TIMSS/PIRLS 4th Grade scale, and Panel B does the same on the PISA scale. Column 1 reports the estimated counterfactual growth in the treatment group, and column 2 reports the ITT effect. Standard errors are clustered at the matched-triplet level. Standard errors are shown in parentheses, and q-values are shown in curly brackets, using the Benjamini and Hochberg (1995) correction for multiple hypothesis testing (MHT), and following Vivalt et al. (2024). Main family measures are highlighted in bold font. * p < 0.10, ** p < 0.05, *** p < 0.01, before adjustment for MHT.

Table A3: *Exploring conditional average treatment effects on student dropout*

	CATE	Standard error	MHT q-value	Weak group	Strong group	Difference
	(1)	(2)	(3)	(4)	(5)	(6)
ITT Effect				-0.001	-0.018	-0.017
Panel A: Student Characteristics						
Male	-0.012	0.002	0.000	38.431	64.650	26.220
Female	-0.005	0.002	0.008	61.569	35.350	-26.220
High Risk of dropout	-0.028	0.007	0.000	4.121	36.741	32.620
Low Risk of dropout	-0.004	0.001	0.015	95.879	63.259	-32.620
Grade 7	-0.005	0.002	0.015	41.191	20.647	-20.544
Grade 8	-0.009	0.002	0.000	29.190	25.157	-4.033
Grade 9	-0.012	0.003	0.000	29.620	54.196	24.576
Panel B: School Characteristics						
Rural	-0.010	0.003	0.003	28.542	29.843	1.301
Urban	-0.008	0.002	0.001	71.458	70.157	-1.301
Regional Development - Low	-0.005	0.002	0.040	50.036	39.977	-10.059
Regional Development - High	-0.012	0.003	0.000	49.964	60.023	10.059
Female students percentage: Bottom quartile	-0.007	0.003	0.034	26.028	19.063	-6.965
Female students percentage: Top quartile	-0.009	0.003	0.015	23.797	26.441	2.645
Number of students: Bottom quartile	-0.006	0.004	0.119	14.435	9.221	-5.214
Number of students: Top quartile	-0.010	0.003	0.003	33.958	46.332	12.374
Number of teachers: Bottom quartile	-0.006	0.004	0.104	18.849	11.358	-7.492
Number of teachers: Top quartile	-0.006	0.003	0.053	34.590	41.698	7.108
Average primary school passing exam score: Bottom quartile	-0.007	0.003	0.034	24.070	22.673	-1.397
Average primary school passing exam score: Top quartile	-0.013	0.004	0.001	21.230	26.730	5.501
Average lower secondary school passing exam score: Bottom quartile	-0.011	0.003	0.002	24.668	29.331	4.663
Average lower secondary school passing exam score: Top quartile	-0.008	0.003	0.018	21.893	21.597	-0.296

Notes. This table explores heterogeneity in treatment effects, with a focus on the average ITT effect on student dropout. The sample consists of all students present in the study schools. The Rank-Weighted Average Treatment Effect (RATE) serves as a formal test for the presence of heterogeneity in treatment effects (Yadlowsky et al., 2025), which in this case is 0.003 (se=0.001). Column 1 reports the conditional average treatment effect (CATE) for each subgroup (defined by the row header), column 2 reports its standard error clustered at the matched-triplet level, and column 3 reports the adjusted q-value, using the Benjamini and Hochberg (1995) correction for multiple hypothesis testing (MHT), and following Vivalt et al. (2024). "Strong group" refers to subgroups whose conditional average treatment effect (CATE) is above the median of all CATEs when switching to the treatment (and below the median for the "Weak group"). A positive number in the Difference column indicates that the average covariate value for the "Strong group" is higher. Regional development refers to a dummy variable indicating any of the following regions: Casablanca-Settat, Fès-Meknès, Marrakech-Safi, Rabat-Salé-Kénitra, or Tanger-Tetouan-Al Hoceima (vs. being in any of the remaining seven regions). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A4: *Exploring conditional average treatment effects on student learning*

	CATE	Standard error	MHT q-value	Weak group	Strong group	Difference
	(1)	(2)	(3)	(4)	(5)	(6)
Group-average ITT Effect				0.21	0.28	0.07
Panel A: Student Characteristics						
Male	0.23	0.03	0.000	48.69	45.96	-2.73
Female	0.26	0.04	0.000	51.31	54.04	2.73
High Risk of dropout	0.29	0.04	0.000	11.32	18.01	6.69
Low Risk of dropout	0.24	0.03	0.000	88.68	81.99	-6.69
Grade 7	0.25	0.04	0.000	30.32	34.20	3.88
Grade 8	0.27	0.04	0.000	30.59	36.90	6.31
Grade 9	0.22	0.04	0.000	39.09	28.91	-10.19
Primary school passing exam score: Bottom quartile	0.26	0.04	0.000	20.99	29.11	8.12
Primary school passing exam score: Top quartile	0.26	0.04	0.000	28.96	20.87	-8.09
Baseline well-being score: Bottom quartile	0.24	0.04	0.000	23.80	26.21	2.41
Baseline well-being score: Top quartile	0.25	0.04	0.000	24.13	25.85	1.71
Baseline test score: Bottom quartile	0.29	0.04	0.000	17.73	32.28	14.55
Baseline test score: Top quartile	0.20	0.04	0.000	35.52	14.46	-21.06
Baseline socioemotional score: Bottom quartile	0.30	0.04	0.000	16.19	33.82	17.63
Baseline socioemotional score: Top quartile	0.22	0.04	0.000	29.76	20.22	-9.54
Baseline creativity score: Bottom quartile	0.21	0.04	0.000	29.87	20.16	-9.72
Baseline creativity score: Top quartile	0.23	0.05	0.000	26.99	22.99	-4.00
Panel B: School Characteristics						
Rural	0.29	0.04	0.000	28.18	42.25	14.08
Urban	0.22	0.03	0.000	71.82	57.75	-14.08
Regional Development - Low	0.21	0.05	0.000	58.12	35.27	-22.85
Regional Development - High	0.28	0.03	0.000	41.88	64.73	22.85
Female students percentage: Bottom quartile	0.25	0.04	0.000	26.66	24.26	-2.39
Female students percentage: Top quartile	0.27	0.07	0.000	21.19	28.28	7.09
Number of students: Bottom quartile	0.23	0.07	0.001	28.76	22.12	-6.64
Number of students: Top quartile	0.22	0.05	0.000	26.03	23.04	-3.00
Number of teachers: Bottom quartile	0.20	0.06	0.001	33.02	23.71	-9.32
Number of teachers: Top quartile	0.21	0.05	0.000	27.22	21.81	-5.41
Average primary school passing exam score: Bottom quartile	0.26	0.06	0.000	20.36	29.71	9.35
Average primary school passing exam score: Top quartile	0.21	0.05	0.000	28.15	21.12	-7.04
Average lower secondary school passing exam score: Bottom quartile	0.26	0.06	0.000	22.66	28.21	5.55
Average lower secondary school passing exam score: Top quartile	0.24	0.06	0.000	25.30	24.53	-0.76

Notes. This table explores heterogeneity in treatment effects, with a focus on the average ITT effect on student learning (Arabic and French). The sample consists of non-attriting 8,959 students who sat the written exams in Arabic or French. The Rank-Weighted Average Treatment Effect (RATE) serves as a formal test for the presence of heterogeneity in treatment effects (Yadlowsky et al., 2025), which in this case is 0.475 (se=0.024). Column 1 reports the conditional average treatment effect (CATE) for each subgroup (defined by the row header), column 2 reports its standard error clustered at the matched-triplet level, and column 3 reports the adjusted q-value, using the Benjamini and Hochberg (1995) correction for multiple hypothesis testing (MHT), and following Vivalt et al. (2024). "Strong group" refers to subgroups whose conditional average treatment effect (CATE) is above the median of all CATEs when switching to the treatment (and below the median for the "Weak group"). A positive number in the Difference column indicates that the average covariate value for the "Strong group" is higher. Regional development refers to a dummy variable indicating any of the following regions: Casablanca-Settat, Fès-Meknès, Marrakech-Safi, Rabat-Salé-Kénitra, or Tanger-Tetouan-Al Hoceima (vs. being in any of the remaining seven regions). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A5: *Exploring conditional average treatment effects on socioemotional skills*

	CATE	Standard error	MHT q-value	Weak group	Strong group	Difference
	(1)	(2)	(3)	(4)	(5)	(6)
Group-average ITT Effect				0.09	0.22	0.12
Panel A: Student Characteristics						
Male	0.18	0.04	0.000	41.13	53.53	12.40
Female	0.13	0.04	0.003	58.87	46.47	-12.40
High Risk of dropout	0.24	0.07	0.001	11.16	18.17	7.01
Low Risk of dropout	0.14	0.04	0.001	88.84	81.83	-7.01
Grade 7	0.20	0.06	0.001	23.26	41.25	17.99
Grade 8	0.19	0.05	0.000	26.64	40.85	14.21
Grade 9	0.07	0.05	0.176	50.10	17.90	-32.20
Primary school passing exam score: Bottom quartile	0.17	0.05	0.002	22.15	27.95	5.80
Primary school passing exam score: Top quartile	0.15	0.05	0.007	26.66	23.17	-3.49
Baseline well-being score: Bottom quartile	0.20	0.05	0.000	21.95	28.06	6.11
Baseline well-being score: Top quartile	0.18	0.05	0.000	26.12	23.86	-2.26
Baseline test score: Bottom quartile	0.18	0.06	0.003	18.58	31.43	12.85
Baseline test score: Top quartile	0.12	0.06	0.039	30.07	19.91	-10.16
Baseline socioemotional score: Bottom quartile	0.24	0.05	0.000	17.64	32.37	14.73
Baseline socioemotional score: Top quartile	0.03	0.06	0.610	38.56	11.43	-27.13
Baseline creativity score: Bottom quartile	0.13	0.06	0.029	26.64	23.39	-3.24
Baseline creativity score: Top quartile	0.16	0.06	0.008	24.09	25.89	1.80
Panel B: School Characteristics						
Rural	0.20	0.06	0.002	31.23	39.20	7.96
Urban	0.13	0.04	0.004	68.77	60.80	-7.96
Regional Development - Low	0.18	0.05	0.001	43.11	50.27	7.16
Regional Development - High	0.13	0.05	0.021	56.89	49.73	-7.16
Female students percentage: Bottom quartile	0.24	0.06	0.000	19.00	31.92	12.92
Female students percentage: Top quartile	0.08	0.07	0.231	31.10	18.37	-12.73
Number of students: Bottom quartile	0.13	0.06	0.044	26.21	24.67	-1.55
Number of students: Top quartile	0.25	0.06	0.000	21.57	27.50	5.93
Number of teachers: Bottom quartile	0.17	0.07	0.011	27.95	28.77	0.82
Number of teachers: Top quartile	0.18	0.06	0.003	24.02	25.00	0.98
Average primary school passing exam score: Bottom quartile	0.17	0.06	0.008	24.27	25.80	1.53
Average primary school passing exam score: Top quartile	0.18	0.07	0.021	24.72	24.55	-0.16
Average lower secondary school passing exam score: Bottom quartile	0.08	0.06	0.231	33.89	16.99	-16.90
Average lower secondary school passing exam score: Top quartile	0.24	0.06	0.001	17.28	32.54	15.26

Notes. This table explores heterogeneity in treatment effects, with a focus on the average ITT effect on student socioemotional skills (proximal outcome measure). The sample consists of non-attriting 8,959 students who sat the written exams in Arabic or French. The Rank-Weighted Average Treatment Effect (RATE) serves as a formal test for the presence of heterogeneity in treatment effects (Yadlowsky et al., 2025), which in this case is 0.666 (se=0.029). Column 1 reports the conditional average treatment effect (CATE) for each subgroup (defined by the row header), column 2 reports its standard error clustered at the matched-triplet level, and column 3 reports the adjusted q-value, using the Benjamini and Hochberg (1995) correction for multiple hypothesis testing (MHT), and following Vivalt et al. (2024). "Strong group" refers to subgroups whose conditional average treatment effect (CATE) is above the median of all CATEs when switching to the treatment (and below the median for the "Weak group"). A positive number in the Difference column indicates that the average covariate value for the "Strong group" is higher. Regional development refers to a dummy variable indicating any of the following regions: Casablanca-Settat, Fès-Meknès, Marrakech-Safi, Rabat-Salé-Kénitra, or Tanger-Tetouan-Al Hoceima (vs. being in any of the remaining seven regions). * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A6: *Exploring conditional average treatment effects on creativity*

	CATE	Standard error	MHT q-value	Weak group	Strong group	Difference
	(1)	(2)	(3)	(4)	(5)	(6)
Group-average ITT Effect				-0.06	0.10	0.16
Panel A: Student Characteristics						
Male	0.05	0.06	0.852	43.45	51.21	7.76
Female	-0.01	0.06	0.899	56.55	48.79	-7.76
High Risk of dropout	0.18	0.08	0.284	10.11	19.22	9.10
Low Risk of dropout	-0.01	0.05	0.899	89.89	80.78	-9.10
Grade 7	0.06	0.07	0.833	26.66	37.86	11.20
Grade 8	0.02	0.07	0.899	32.64	34.84	2.20
Grade 9	-0.03	0.07	0.899	40.70	27.30	-13.40
Primary school passing exam score: Bottom quartile	0.09	0.07	0.650	20.27	29.82	9.55
Primary school passing exam score: Top quartile	0.08	0.06	0.650	26.81	23.01	-3.80
Baseline well-being score: Bottom quartile	-0.01	0.07	0.899	24.83	25.18	0.35
Baseline well-being score: Top quartile	0.03	0.07	0.899	25.47	24.51	-0.97
Baseline test score: Bottom quartile	0.09	0.07	0.650	16.92	33.08	16.16
Baseline test score: Top quartile	-0.03	0.06	0.899	30.59	19.40	-11.19
Baseline socioemotional score: Bottom quartile	0.03	0.07	0.899	23.76	26.25	2.49
Baseline socioemotional score: Top quartile	-0.01	0.07	0.899	27.22	22.77	-4.45
Baseline creativity score: Bottom quartile	0.16	0.08	0.334	13.15	36.88	23.72
Baseline creativity score: Top quartile	-0.15	0.09	0.505	39.81	10.18	-29.63
Panel B: School Characteristics						
Rural	0.05	0.10	0.993	28.22	42.21	13.99
Urban	0.00	0.06	0.993	71.78	57.79	-13.99
Regional Development - Low	0.04	0.07	0.993	43.16	50.22	7.07
Regional Development - High	-0.00	0.07	0.993	56.84	49.78	-7.07
Female students percentage: Bottom quartile	0.11	0.09	0.993	20.67	30.25	9.57
Female students percentage: Top quartile	0.09	0.11	0.993	18.53	30.94	12.41
Number of students: Bottom quartile	0.04	0.09	0.993	23.15	27.72	4.57
Number of students: Top quartile	0.11	0.12	0.993	22.93	26.14	3.21
Number of teachers: Bottom quartile	-0.01	0.09	0.993	28.33	28.39	0.06
Number of teachers: Top quartile	0.04	0.10	0.993	24.56	24.46	-0.09
Average primary school passing exam score: Bottom quartile	0.04	0.12	0.993	20.34	29.73	9.39
Average primary school passing exam score: Top quartile	0.04	0.10	0.993	24.69	24.58	-0.12
Average lower secondary school passing exam score: Bottom quartile	0.01	0.08	0.993	25.65	25.22	-0.43
Average lower secondary school passing exam score: Top quartile	0.15	0.12	0.993	18.46	31.36	12.90

Notes. This table explores heterogeneity in treatment effects, with a focus on the average ITT effect on creativity. The sample consists of non-attriting 8,959 students who sat the written exams in Arabic or French. The Rank-Weighted Average Treatment Effect (RATE) serves as a formal test for the presence of heterogeneity in treatment effects (Yadlowsky et al., 2025), which in this case is 0.699 (se=0.038). Column 1 reports the conditional average treatment effect (CATE) for each subgroup (defined by the row header), column 2 reports its standard error clustered at the matched-triplet level, and column 3 reports the adjusted q-value, using the Benjamini and Hochberg (1995) correction for multiple hypothesis testing (MHT), and following Vivalt et al. (2024). "Strong group" refers to subgroups whose conditional average treatment effect (CATE) is above the median of all CATEs when switching to the treatment (and below the median for the "Weak group"). A positive number in the Difference column indicates that the average covariate value for the "Strong group" is higher. Regional development refers to a dummy variable indicating any of the following regions: Casablanca-Settat, Fès-Meknès, Marrakech-Safi, Rabat-Salé-Kénitra, or Tanger-Tetouan-Al Hoceima (vs. being in any of the remaining seven regions). * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A7: *Intent-to-treat effects on subdomains of learning*

	Treatment group	Main results
	Counterfactual growth	Overall ITT effect
	(1)	(2)
Panel A: Arabic		
Written	0.34***	0.20***
	(0.04)	(0.04)
		{0.000}
Reading and Comprehension	0.31***	0.22***
	(0.03)	(0.03)
		{0.000}
Written Production	0.26***	0.12*
	(0.06)	(0.06)
		{0.051}
Oral	0.04	0.31***
	(0.05)	(0.05)
		{0.000}
Panel B: French		
Written	0.21***	0.29***
	(0.05)	(0.05)
		{0.000}
Reading and Comprehension	0.18***	0.39***
	(0.05)	(0.05)
		{0.000}
Written Production	0.16***	0.19***
	(0.05)	(0.05)
		{0.000}
Oral	0.23***	0.35***
	(0.05)	(0.05)
		{0.000}
Panel C: Math		
Below-level	0.16***	0.19***
	(0.04)	(0.04)
		{0.000}
At-level	0.16***	0.19***
	(0.04)	(0.04)
		{0.000}
Applying or Reasoning	0.18***	0.23***
	(0.05)	(0.05)
		{0.000}
Knowing	0.29***	0.20***
	(0.06)	(0.06)
		{0.001}
Algebra	0.21***	0.30***
	(0.04)	(0.04)
		{0.000}
Data and Chance	0.45***	0.34***
	(0.06)	(0.06)
		{0.000}
Geometry	0.38***	0.28***
	(0.07)	(0.07)
		{0.000}
Number	-0.20***	0.46***
	(0.05)	(0.05)
		{0.000}
Panel D: Science		
Applying or Reasoning	0.87***	1.19***
	(0.06)	(0.06)
		{0.000}
Knowing	-0.91***	1.68***
	(0.11)	(0.11)
		{0.000}
Biology	0.79***	1.15***
	(0.07)	(0.07)
		{0.000}
Physics	0.00	1.24***
	(0.06)	(0.06)
		{0.000}
Chemistry		2.31***
		(0.08)
		{0.000}

Notes. This table reports on the program's intent-to-treat (ITT) effects following equation 1, on students' learning among the 18,512 non-attriting students across our 300 government lower secondary schools. Column 1 reports the estimated counterfactual growth in the treatment group, and column 2 reports the ITT effect. Standard errors are clustered at the matched-triplet level. "Chemistry" was not assessed at baseline, hence it does not have a counterfactual growth and is estimated using the ANCOVA specification in equation 2. Standard errors are clustered at the matched-triplet level in column 2. Standard deviations are shown in brackets; standard errors are shown in parentheses. *q*-values are shown in curly brackets, using the Benjamini and Hochberg (1995) correction for multiple hypothesis testing (MHT), and following Vivalt et al. (2024). Main family measures are highlighted in bold font. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, before adjustment for MHT.

Table A8: *Effect on subdomains of socioemotional skills and creativity*

		By Language	
	Overall	Arabic	French
	(1)	(2)	(3)
Proximal outcome measure	0.22***	0.26***	0.22***
	(0.03)	(0.05)	(0.04)
	{0.000}	{0.000}	{0.000}
Interpersonal skills	0.14***	0.18***	0.14**
	(0.04)	(0.06)	(0.05)
	{0.001}	{0.002}	{0.041}
Pro-sociality	0.20***	0.23***	0.20***
	(0.04)	(0.05)	(0.04)
	{0.000}	{0.000}	{0.001}
Emotion perception	-0.03	-0.03	-0.03
	(0.03)	(0.04)	(0.04)
	{0.315}	{0.431}	{0.538}
Intrapersonal skills	0.26***	0.31***	0.26***
	(0.04)	(0.05)	(0.05)
	{0.000}	{0.000}	{0.001}
Perceived control	0.28***	0.33***	0.28***
	(0.04)	(0.05)	(0.05)
	{0.000}	{0.000}	{0.001}
Self-regulation and discipline	0.16***	0.20***	0.16**
	(0.04)	(0.05)	(0.05)
	{0.000}	{0.000}	{0.028}
Growth mindset	0.20***	0.21***	0.20***
	(0.03)	(0.04)	(0.05)
	{0.000}	{0.000}	{0.001}
Locus of control	0.45***	0.54***	0.45***
	(0.04)	(0.05)	(0.05)
	{0.000}	{0.000}	{0.000}
Self-efficacy	0.11***	0.16***	0.11
	(0.04)	(0.05)	(0.05)
	{0.003}	{0.001}	{0.202}
Grit	0.18***	0.23***	0.18***
	(0.03)	(0.04)	(0.04)
	{0.000}	{0.000}	{0.003}
Work discipline and diligence	0.18***	0.23***	0.18***
	(0.03)	(0.04)	(0.04)
	{0.000}	{0.000}	{0.002}
Creativity	0.05	0.01	0.05
	(0.05)	(0.06)	(0.06)
	{0.340}	{0.910}	{0.155}

Notes. This table reports on the program's intent-to-treat (ITT) effects following equation 1 on students' socioemotional skills and creativity among the 8,959 non-attriting students across our 300 government lower secondary schools who sat the written exams in Arabic or French. Column (1) reports the overall ITT effect, and columns (2) and (3) respectively report the effects separately for the two subsamples of students (students who sat the written exams in Arabic or French). Standard errors are clustered at the matched-triplet level. "Pro-sociality" was not assessed at baseline, hence it is estimated using the ANCOVA specification in equation 2. Standard errors are clustered at the matched-triplet level in columns 1, 2, and 3. Standard deviations are shown in brackets; standard errors are shown in parentheses. *q*-values are shown in curly brackets, using the Benjamini and Hochberg (1995) correction for multiple hypothesis testing (MHT), and following Vivalt et al. (2024). Main family measures are highlighted in bold font. * p < 0.10, ** p < 0.05, *** p < 0.01, before adjustment for MHT.

Table A9: *Program implementation, exposure, and mechanisms (by program type)*

	Without Socioemotional		With Socioemotional	
	Treatment Average	ITT Effect	Treatment Average	ITT Effect
	(1)	(2)	(3)	(4)
Panel A: Program implementation and exposure				
Student received TaRL	0.96	0.95***	0.96	0.95***
	[0.21]	(0.01)	[0.20]	(0.01)
Student received explicit teaching	0.89	0.65***	0.89	0.68***
	[0.31]	(0.02)	[0.31]	(0.03)
Students participation in extra-curricular activities	0.65	0.21***	0.67	0.19***
	[0.48]	(0.02)	[0.47]	(0.02)
Student participated in socioemotional workshops	0.22	0.11***	0.52	0.40***
	[0.41]	(0.02)	[0.50]	(0.02)
Student participated in tutoring program	0.53	-0.02	0.54	-0.04
	[0.50]	(0.02)	[0.50]	(0.03)
At-risk student participated in tutoring program	0.68	0.13***	0.70	0.07
	[0.47]	(0.04)	[0.46]	(0.05)
Targeted student participated in socioemotional workshops	0.22	0.11***	0.66	0.55***
	[0.41]	(0.02)	[0.47]	(0.03)
Panel B: Mechanisms				
School climate and well-being	0.10	0.13***	0.14	0.13**
	[1.05]	(0.05)	[1.08]	(0.05)
Knows social specialist	0.44	0.31***	0.58	0.42***
	[0.50]	(0.04)	[0.49]	(0.04)
Study habits (spent more than 30 min/day doing homework after school)	0.96	0.01	0.96	0.01
	[0.19]	(0.01)	[0.19]	(0.01)

Notes. This table describes the program's implementation and exposure to students in the 300 study schools among our non-attriting sample of 18,512 students. "Without socioemotional" refers to the 116 Pioneer schools that did not receive the socioemotional workshops. "With socioemotional" refers to the 84 Pioneer schools that were assigned to receive the socioemotional workshops. "Treatment Average" reports the average for the specific program type. "ITT Effect" reports on the regression-adjusted difference between pioneer schools (by program type) and comparison schools, controlling for matched-triplet-by-grade fixed effects, in columns 3 and 4. Reversed outcomes were flipped so higher scores represent desirable outcomes. Standard errors are clustered at the matched-triplet in columns 3 and 4. Standard deviations are shown in brackets; standard errors are shown in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01.

Table A10: *Intent-to-treat effects on student dropout and learning (grades 7 and 8)*

	Without Socioemotional		With Socioemotional	
	Counterfactual Levels (A) / Growth (B)	Component ITT Effect	Counterfactual Levels (A) / Growth (B)	Component ITT Effect
	(1)	(2)	(3)	(4)
Panel A: Dropout and repetition				
Dropout by end of school year (global: either excluded or dropout)	0.055	-0.020***	0.055	-0.016***
	[0.003]	(0.003)	[0.003]	(0.003)
		{0.000}		{0.000}
Dropout by end of the school year	0.044	-0.015***	0.044	-0.011***
	[0.003]	(0.003)	[0.003]	(0.003)
		{0.000}		{0.000}
Excluded by end of school year	0.011	-0.005***	0.011	-0.005**
	[0.002]	(0.002)	[0.002]	(0.002)
		{0.003}		{0.003}
Repeated at end of school year	0.217	-0.093***	0.217	-0.102***
	[0.009]	(0.007)	[0.009]	(0.007)
		{0.000}		{0.000}
Long term dropout	0.074	-0.019***	0.074	-0.014***
	[0.004]	(0.003)	[0.004]	(0.003)
		{0.000}		{0.000}
Panel B: Learning				
Overall (stacked test scores)	-0.01	0.55***	0.20***	0.62***
	(0.04)	(0.04)	(0.04)	(0.04)
		{0.000}		{0.000}
Arabic	0.21***	0.23***	0.25***	0.29***
	(0.05)	(0.05)	(0.05)	(0.05)
		{0.000}		{0.000}
French	0.06	0.30***	0.22**	0.31***
	(0.05)	(0.05)	(0.05)	(0.05)
		{0.000}		{0.000}
Math	-0.07	0.28***	-0.02	0.33***
	(0.06)	(0.06)	(0.06)	(0.06)
		{0.000}		{0.000}
Science	-0.15	1.20***	0.35***	1.30***
	(0.08)	(0.08)	(0.08)	(0.08)
		{0.000}		{0.000}
Physics & Chemistry	-0.50***	1.16***	-0.03	1.35***
	(0.08)	(0.08)	(0.08)	(0.08)
		{0.000}		{0.000}
Life Science	0.27**	1.17***	0.76***	1.16***
	(0.09)	(0.09)	(0.09)	(0.09)
		{0.000}		{0.000}

Notes. This table reports on the program's intent-to-treat (ITT) effects following equation 1. "Without socioemotional" refers to the 116 Pioneer schools that did not receive the socioemotional support workshops. "With socioemotional" refers to the 84 Pioneer schools that were assigned to receive the socioemotional support workshops. By program variant, Panel A reports the effect of the program on dropout among all students present in the study schools relative to the dropout in the prior year. Columns 1 and 3 report the estimated counterfactual levels of dropout in treatment schools, and columns 2 and 4 report the ITT effects. Panel B shows the effect of the program on student learning among the 15,466 non-attriting students across our 300 government lower secondary schools. Columns 1 and 3 report the estimated counterfactual growth in learning in treatment schools, and columns 2 and 4 report the ITT effects. Standard errors are clustered at the matched-triplet level. Standard deviations are shown in brackets; standard errors are shown in parentheses. *q*-values are shown in curly brackets, using the Benjamini and Hochberg (1995) correction for multiple hypothesis testing (MHT), and following Vivalt et al. (2024). Main family measures are highlighted in bold font. * p < 0.10, ** p < 0.05, *** p < 0.01, before adjustment for MHT.

Table A11: *Intent-to-treat effects on socioemotional skills and creativity (grades 7 and 8)*

	Without Socioemotional		With Socioemotional	
	Counterfactual Growth	ITT Effect	Counterfactual Growth	ITT Effect
	(1)	(2)	(3)	(4)
Panel A: Socioemotional skills				
Proximal outcome measure	-2.24***	0.28***	-2.06***	0.27***
	(0.05)	(0.05)	(0.05)	(0.05)
		{0.000}		{0.000}
Interpersonal skills	0.18***	0.21***	0.30***	0.21***
	(0.05)	(0.05)	(0.05)	(0.05)
		{0.000}		{0.000}
Pro-sociality		0.23***		0.22***
		(0.04)		(0.04)
		{0.000}		{0.000}
Emotion perception	0.19***	0.01	0.20***	0.04
	(0.04)	(0.04)	(0.04)	(0.04)
		{0.759}		{0.759}
Intrapersonal skills	-2.77***	0.33***	-2.55***	0.23***
	(0.05)	(0.05)	(0.05)	(0.05)
		{0.000}		{0.000}
Perceived control	-2.46***	0.36***	-2.23***	0.29***
	(0.06)	(0.06)	(0.06)	(0.06)
		{0.000}		{0.000}
Self-regulation and discipline	-2.35***	0.20***	-2.20***	0.11
	(0.05)	(0.05)	(0.05)	(0.05)
		{0.000}		{0.000}
Panel B: Creativity				
Creativity	0.12*	0.16**	0.28***	-0.06
	(0.06)	(0.06)	(0.06)	(0.06)
		{0.012}		{0.012}

Notes. This table reports on the program's intent-to-treat (ITT) effects following equation 1 on students' socioemotional skills and creativity among the 5,913 non-attriting students across our 300 government lower secondary schools who sat the written exams in Arabic or French. "Without socioemotional" refers to the 116 Pioneer schools that did not receive the socioemotional support intervention. "With socioemotional" refers to the 84 Pioneer schools that were assigned to receive the socioemotional support intervention. Columns 1 and 3 report the estimated counterfactual growth in the treatment group, and columns 2 and 4 report the respective ITT effect for the respective program types. Standard errors are clustered at the matched-triplet level. "Prosociality" was not assessed at baseline, hence it does not have a counterfactual growth and is estimated using the ANCOVA specification in equation 2. Standard deviations are shown in brackets; standard errors are shown in parentheses. *q*-values are shown in curly brackets, using the Benjamini and Hochberg (1995) correction for multiple hypothesis testing (MHT), and following Vivalt et al. (2024). Main family measures are highlighted in bold font. * p < 0.10, ** p < 0.05, *** p < 0.01, before adjustment for MHT.

Table A12: *Sensitivity to differential attrition*

	Learning (stacked test scores)	Socioemotional skills (proximal outcome measure)	Creativity
	(1)	(2)	(3)
Panel A: Lee (2009) bounds			
Lower bound			
Treatment	0.38***	0.04	-0.09*
	(0.03)	(0.03)	(0.05)
	{0.000}	{0.239}	{0.101}
Upper bound			
Treatment	0.59***	0.40***	0.20***
	(0.03)	(0.03)	(0.05)
	{0.000}	{0.000}	{0.000}
Panel B: Behaghel et al. (2015) bounds: Attrition based on number of days tracked			
Lower bound			
Treatment	0.50***	0.21***	0.03
	(0.03)	(0.03)	(0.05)
	{0.000}	{0.000}	{0.515}
Upper bound			
Treatment	0.53***	0.24***	0.06
	(0.03)	(0.03)	(0.05)
	{0.000}	{0.000}	{0.252}
Panel C: IPW - Attrition based on baseline covariates			
Treatment	0.52***	0.22***	0.05
	(0.03)	(0.03)	(0.05)
	{0.000}	{0.000}	{0.340}
Panel D: IPW - Attrition based on baseline covariates and number of days tracked			
Treatment	0.54***	0.23***	0.04
	(0.03)	(0.03)	(0.05)
	{0.000}	{0.000}	{0.340}

Notes. Estimated treatment effects. The dependent variables in columns 1, 2 and 3 are the family indices from Tables 3 and 4. Treatment captures effects for schools which received the treatment. Panel A estimates bounds for the ITT estimates following Lee (2009). Panel B estimates tightened bounds for the ITT estimates following Behaghel et al. (2015). Panel C estimates an IPW estimate for the ITT effect using baseline covariates as predictors of attrition. Panel D estimates an IPW estimate for the ITT effect using baseline covariates as predictors of attrition following Molina-Millán and Macours (2025). Probability of attrition is estimated through a probit regression of an attrition indicator on treatment status and baseline covariates in panels A, B, C, and on treatment status, baseline covariates and number of days tracked in panel D. "Number of days tracked" information includes the number of days visits to a school before a student was surveyed, and its quadratic. All regressions follow the specifications in tables 3 and 4 except for the weights. Standard errors are clustered at the matched-triplet level and are shown in parentheses. q-values are shown in curly brackets, using the Benjamini and Hochberg (1995) correction for multiple hypothesis testing (MHT), and following Vivalt et al. (2024). Main family measures are highlighted in bold font. * p < 0.10, ** p < 0.05, *** p < 0.01, before adjustment for MHT.

Table A13: *Effect on socioemotional skills and creativity: Robustness to social desirability bias*

		Social desirability	
	Difference	Below median	Above median
	(1)	(2)	(3)
Proximal outcome measure	0.02	0.23***	0.25***
	(0.03)	(0.04)	(0.04)
		{0.000}	{0.000}
Interpersonal skills	0.13***	0.10**	0.23***
	(0.04)	(0.05)	(0.05)
		{0.025}	{0.000}
Pro-sociality	-0.01	0.21***	0.21***
	(0.05)	(0.04)	(0.05)
		{0.000}	{0.000}
Emotion perception	0.04	-0.03	0.00
	(0.03)	(0.03)	(0.04)
		{0.295}	{0.965}
Intrapersonal skills	-0.16***	0.33***	0.17***
	(0.03)	(0.04)	(0.04)
		{0.000}	{0.000}
Perceived control	-0.12***	0.34***	0.22***
	(0.04)	(0.04)	(0.05)
		{0.000}	{0.000}
Self-regulation and discipline	-0.17***	0.23***	0.07
	(0.03)	(0.04)	(0.04)
		{0.000}	{0.132}
Growth mindset	0.01	0.21***	0.22***
	(0.03)	(0.04)	(0.04)
		{0.000}	{0.000}
Locus of control	0.14***	0.39***	0.53***
	(0.03)	(0.04)	(0.04)
		{0.000}	{0.000}
Self-efficacy	-0.04	0.13***	0.09**
	(0.03)	(0.04)	(0.04)
		{0.001}	{0.044}
Grit	0.13***	0.14***	0.27***
	(0.03)	(0.04)	(0.04)
		{0.000}	{0.000}
Work discipline and diligence	-0.23***	0.28***	0.05
	(0.03)	(0.03)	(0.04)
		{0.000}	{0.193}
Creativity	-0.06*	0.05	-0.01
	(0.03)	(0.05)	(0.06)
		{0.290}	{0.896}

Notes. This table reports on the heterogeneity in the program's intent-to-treat (ITT) effects following equation 1 on students' socioemotional skills and creativity among the 8,498 non-attriting students across our 300 government lower secondary schools who have non-missing social desirability scores at baseline and who sat the written exams in Arabic or French. Column (1) reports the difference in the effect on students who exhibit high social desirability bias and students who exhibit low social desirability bias, and columns (2) and (3) respectively report the effects separately for the two subsamples of students. Standard errors are clustered at the matched-triplet level. "Pro-sociality" was not assessed at baseline, hence it is estimated using the ANCOVA specification in equation 2. Standard errors are clustered at the matched-triplet level in columns 1, 2, and 3. Standard deviations are shown in brackets; standard errors are shown in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A14: *Intent-to-treat effects on student transfers at/before the start of the academic year*

	Counterfactual Level	Main results ITT effect
	(1)	(2)
Panel A: Transferring out between		
July and August	0.158	0.020*
	[0.012]	(0.010)
July and September	0.161	0.020*
	[0.012]	(0.011)
July and October	0.229	0.020
	[0.014]	(0.013)
August and September	0.001	0.001
	[0.003]	(0.003)
August and October	0.059	0.006
	[0.008]	(0.006)
September and October	0.057	0.005
	[0.006]	(0.005)
Panel B: Transferring in between		
July and August	0.245	0.005
	[0.009]	(0.008)
July and September	0.254	0.004
	[0.009]	(0.008)
July and October	0.321	-0.001
	[0.009]	(0.008)
August and September	0.016	-0.006
	[0.009]	(0.008)
August and October	0.090	-0.018*
	[0.011]	(0.010)
September and October	0.074	-0.011**
	[0.006]	(0.005)

Notes. This table reports on the program's intent-to-treat (ITT) effects following equation 1. Column 1 reports the estimated counterfactual level of transfers in treatment schools, and column 2 reports the ITT effect. Panel A reports the effect of the program on students transferring out of a treatment school relative to the prior year. Panel B reports the effect of the program on students transferring into a treatment school relative to the prior year. Standard errors are clustered at the matched-triplet level. Standard deviations are shown in brackets; standard errors are shown in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A15: *Robustness of intent-to-treat effects on student dropout to alternative sample definitions*

	Counterfactual Level	Main results ITT effect
	(1)	(2)
Panel A: Students enrolled at the start of July		
Dropout by end of school year (global: either excluded or dropout)	0.060	-0.019***
	[0.004]	(0.003)
Dropout by end of the school year	0.043	-0.014***
	[0.002]	(0.002)
Excluded by end of school year	0.017	-0.005*
	[0.003]	(0.003)
Repeated at end of school year	0.195	-0.076***
	[0.008]	(0.007)
Long term dropout	0.061	-0.017***
	[0.003]	(0.002)
Panel B: Students enrolled at the start of August		
Dropout by end of school year (global: either excluded or dropout)	0.056	-0.018***
	[0.003]	(0.003)
Dropout by end of the school year	0.041	-0.014***
	[0.002]	(0.002)
Excluded by end of school year	0.015	-0.004*
	[0.002]	(0.002)
Repeated at end of school year	0.217	-0.097***
	[0.008]	(0.006)
Long term dropout	0.061	-0.016***
	[0.003]	(0.002)
Panel C: Students enrolled at the start of October		
Dropout by end of school year (global: either excluded or dropout)	0.057	-0.020***
	[0.003]	(0.003)
Dropout by end of the school year	0.042	-0.015***
	[0.002]	(0.002)
Excluded by end of school year	0.015	-0.005**
	[0.002]	(0.002)
Repeated at end of school year	0.223	-0.104***
	[0.008]	(0.006)
Long term dropout	0.061	-0.017***
	[0.003]	(0.002)

Notes. This table reports on the program's intent-to-treat (ITT) effects following equation 1. Column 1 reports the estimated counterfactual level of dropout in treatment schools, and column 2 reports the ITT effect. Panel A reports the effect of the program on students who were reported as enrolled at the start of July (when the previous academic year ends). Panel B reports the effect of the program on students who were reported as enrolled at the start of August (when the new academic year starts). Panel C reports the effect of the program on students who were reported as enrolled at the start of October (one month into the new academic year). Standard errors are clustered at the matched-triplet level. Standard deviations are shown in brackets; standard errors are shown in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

B Measurement

B.1 Psychometric properties

Table B1 provides psychometric properties of the academic 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).[30]

In Arabic, French, math, and science, students attempted almost all test questions, leaving only 1.02, 4.07, 5.40, and 3.05 percent unanswered (see column 6). This finding suggests that the assessments were of manageable duration and produced limited respondent fatigue among students.

Virtually all items performed well. Hardly any test questions had to be removed due to unfavorable measurement properties (see column 2), and the average test item discriminated very well (see column 8)[31]. In addition, the tests proved to be internally consistent, with average item-test correlations ranging from 0.28 to 0.38. Despite our efforts to include many easy items, the academic skill assessments were difficult for students, which is reflected by a low overall percentage of correctly answered test questions (see column 4) and the average difficulty parameter (see column 9).

The average conditional reliability, reported in column (10) of B1, measures the precision of each instrument across the spectrum of respondent abilities.[32] 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 high levels of precision. This is true for a wide 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 are generally lower.

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

[31] 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.

[32] To report on the reliability of our measures, we calculate the average conditional reliability based on the IRT estimates across the sample:

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

where N is the sample size, $\mathrm{Var}(\hat{\theta})$ is the variance of the latent proficiency estimates in the sample, and $\mathrm{SE}(\hat{\theta}_i)$ is the standard error of measurement for student i. This measure reflects the proportion of observed score variance attributable to true differences in ability. For comparison, Cronbach’s alpha (reported in column 6) provides the corresponding Classical Test Theory-based measure of reliability; as expected, the two measures yield highly consistent results.

B.2 Linkage to international assessments

To facilitate international comparisons, we linked our test scores to established international large-scale assessments: the Programme for International Student Assessment (PISA), the Progress in International Reading Literacy Study (PIRLS), and the Trends in International Mathematics and Science Study (TIMSS).

This linking relied on a common-item, non-equivalent groups design, implemented through two-group Item Response Theory (IRT) models. This approach allows our baseline and endline scores to be expressed on the same international scale by leveraging a subset of anchor items that were originally calibrated in the corresponding international sample.

All models were estimated through multi-group IRT models by marginal maximum likelihood with expected-a-posteriori (EAP) scoring. We employed a two-parameter logistic (2PL) specification for dichotomous items and a partial-credit model (PCM) for items with ordered response categories.

We followed the following steps:

1. Test for differential item functioning (DIF) between the international parameters and those estimated in the Moroccan sample.
2. Estimate a single=group IRT model using the baseline observations only, constraining the item parameters for the anchor items.
3. Link baseline to endline by freeing parameters for items exhibiting DIF.
4. Estimate a two-group IRT model on endline data with item parameters constrained to their international values.
5. Rescaling the resulting latent proficiency estimates to the international reporting metric.

This procedure yields proficiency scores directly comparable to international benchmarks (e.g., mean = 500, SD = 100 in PISA and TIMSS), allowing Moroccan student performance to be interpreted within a globally standardized framework.

B.2.1 Differential Item Functioning (DIF) and validation of anchors

As a preliminary step, we assessed the invariance of international anchor items within the Moroccan context. A critical assumption for cross-national linking is that the relative difficulty of items remains stable across populations. We estimate a multi-group Item Response Theory (IRT) model, pooling our sample with international item parameters[33].

We then conducted a series of likelihood-ratio tests (LRTs) to examine invariance item-by-item. In each iteration, we freed the parameters of a single anchor item and compared the fit of this partially freed model to the fully constrained baseline. Items whose release produced a statistically significant improvement in model fit were flagged for DIF and subsequently excluded from the anchor set. This procedure ensured that only items exhibiting stable cross-population properties served to place our study's ability estimates on the international metric.

We further evaluated uniform versus non-uniform DIF by comparing Test Characteristic Curves (TCCs) across three models: (i) all anchor parameters freely estimated, (ii) all anchors constrained to their international values, and (iii) only the subset of anchors retained after our LRT-based screening. As shown in Figure B1, the TCCs for the retained anchors align closely across specifications, suggesting that any residual DIF is uniform and unlikely to introduce systematic bias in the latent proficiency estimates.

B.2.2 Calibration and linking

To establish an initial benchmark, we estimated a single-group IRT model using only the baseline observations. In this stage, the discrimination (α_j) and difficulty (β_j) parameters for the subset of international anchor items retained from Step 1 were fixed to their official international values. The parameters for all non-anchor items (Morocco-specific items) were freely estimated. This step places our baseline anchors on the international metric without imposing assumptions about the distribution of latent proficiency in the Moroccan sample.

To ensure comparability across assessment waves, we linked the baseline and endline scales using a common-item, non-equivalent groups design. Specifically, we appended the baseline observations to the endline control group and estimated a two-group IRT model in which we constrained the item parameters of anchors. The endline control group served as the reference metric, preserving the international scale at endline while allowing baseline means and variances to adjust freely. This procedure places both baseline and endline scores on a consistent scale.

In the final calibration step, we estimated a two-group IRT model using the full endline sample, treating students in pioneer and non-pioneer schools as separate groups. While their latent means and variances were allowed to vary by treatment status to capture the

[33] Anchor items were initially fixed to their international discrimination and difficulty values.

intervention's impact, the item parameters remained constrained to the values established in the previous steps.

Specifically, the international anchors were fixed to their original international parameters, and the Morocco-specific items were fixed to the values derived in the linking step. This produced internationally comparable endline scores for both treatment and control groups, while preserving potential shifts in the latent distribution attributable to the intervention.

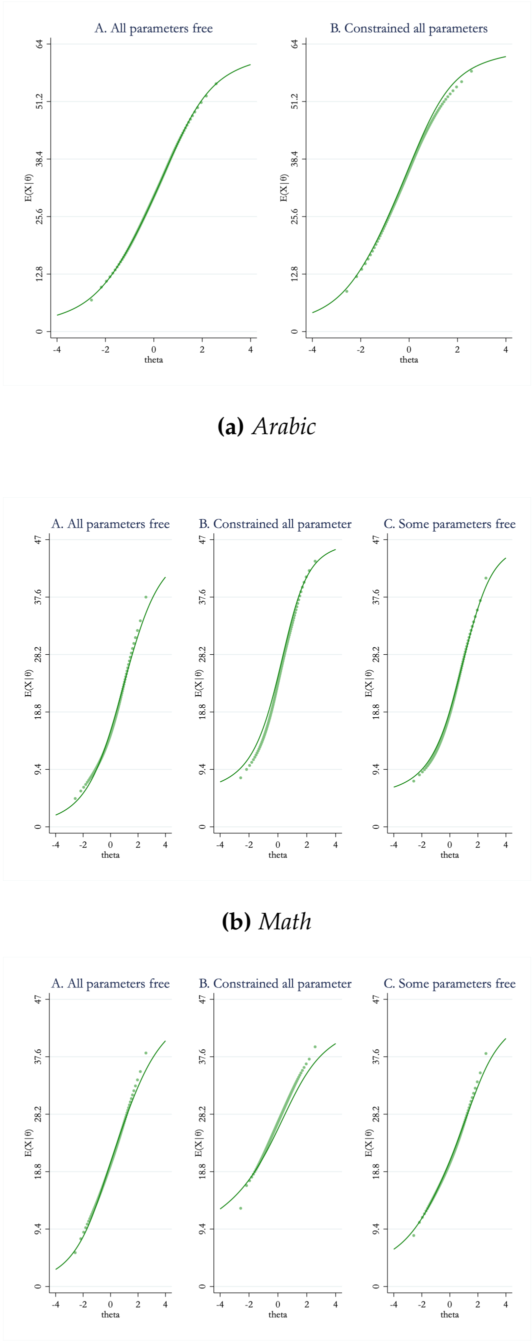
B.2.3 Transformation to the international reporting metric

Finally, we transformed the ability estimates, θ_i, using the standard international reporting scale using a linear transformation:

$$Score_i = 500 + 100(\theta) \tag{3}$$

For each assessment, because the scale is anchored on the endline control distribution, this transformation preserves comparability over time and across treatment conditions while maintaining alignment with the mean (500) and standard deviation (100) in the international population.

Figure B1: *Test Characteristic Curves*



(a) *Arabic*

(b) *Math*

(c) *Science*

Notes: The figure reports test characteristic curves (TCCs) under alternative parameter constraints. Panels A allow all item parameters to vary freely. Panels B constrain international anchor item parameters to their official values. Panels C constrain only the subset of international anchor items retained after the differential item functioning (DIF) analysis. Math and Science scores are linked to TIMSS; Arabic scores are linked to PISA. For Arabic, Panel C is omitted because no statistically significant DIF was detected for the PISA anchor items, so it was not required to free up additional item parameters. We do not have an international reference for our French assessment.

Table B1: *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	Marginal reliability
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Assessments										
Arabic	94	1	52	26.02	1.02			1.16	0.19	0.89
French	89	0	36	19.93	4.07			1.22	1.01	0.89
Math	53	0	11	35.43	5.40		0.36	0.89	1.18	0.78
Science	55	1	10	40.32	3.05		0.28	0.72	1.01	0.77
Physics and Chemistry	30	0	5	40.17	2.90		0.33	0.74	1.35	0.62
Life Science	26	0	5	40.30	3.25		0.38	0.85	0.55	0.66
Panel B: Other measures										
Creativity (Torrance)	7	0								
Student Perceived Control	13	0						0.84		0.75
Growth Mindset	3	0						1.91		0.69
Locus of control	5	0						1.09		0.68
Student self-efficacy	5	0						1.18		0.66
Self-discipline index (Self-Regulation, Disc, Diligence)	25	0						0.89		0.88
Work discipline and diligence	17	0						1.00		0.89
Self-regulation (short self-control scale)	8	0						0.85		0.71
Perceiving emotions (PAGE)	16	1						0.53		0.62
Wellbeing index	18	0						0.99		0.69
Feeling of belonging in school	5	0						1.15		0.61
Bullying (reversed)	9	0						1.54		0.58
Perceived stress (PSS-4, reversed)	4	0						0.89		0.57
Grit	8	0						0.92		0.82

Notes. Sample and unit of observation: 18,512 assessed students across the 300 schools in the study. This table reports on the measurement properties of student assessment instruments and survey instruments included in the study's baseline and endline. "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. * p < 0.10, ** p < 0.05, *** p < 0.01.

B.3 Manipulation checks

B.3.1 Comparison of open-ended vs. multiple-choice items

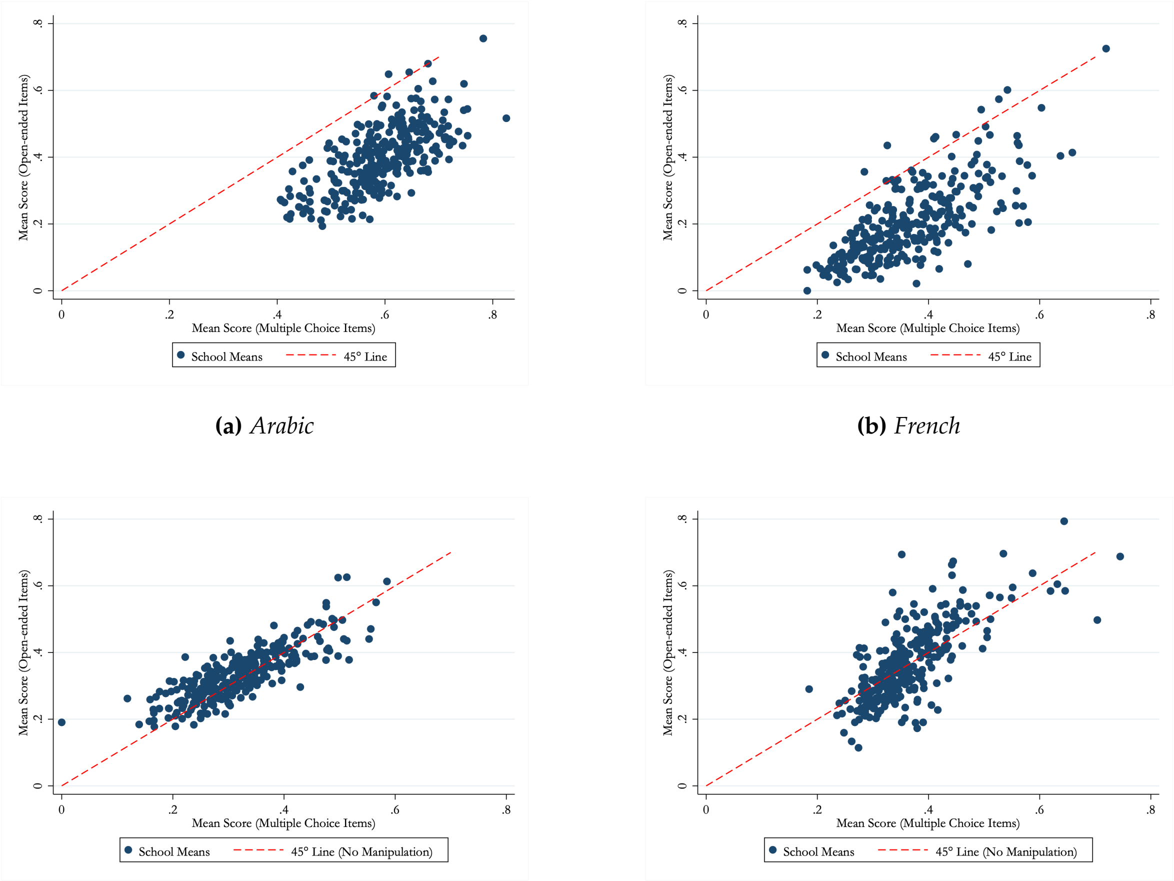
We run a few diagnostics to check for potential score manipulation by comparing open-ended and multiple-choice items at the endline. The idea is that open-ended items—since they require subjective scoring—should typically be harder (lower means) and show greater variability than multiple-choice ones. If we saw the opposite (open-ended items scoring higher or showing compressed variance), that could suggest lenient marking or over-correction.

For each subject, we computed the difference in mean scores (open multiple choice). Positive values could indicate over-scoring on open-ended items. Ratio of standard deviations (SD_{open}/SD_{MC}). Ratios below 1 suggest less dispersion in open-ended scoring.

Additionally, we plotted the school-level means for open-ended items (y-axis) against multiple-choice items (x-axis), for each subject (Arabic, French, Math, and Science). Results look clean for Math and Science: both subjects show small differences in means and similar dispersion across item typesconsistent with objective, well-calibrated scoring. For Arabic and French, patterns are slightly less balanced. In French, the difference in means is 0.16, and the ratio of standard deviations is about 1.11. That means open-ended items are somewhat harder (lower average) and slightly more variable than multiple-choice, so no sign of inflation, but the higher variance could reflect greater subjectivity in scoring. Arabic shows a similar tendency: open-ended items are more dispersed, which could stem from broader differences in teacher interpretation or student writing ability.

Table B2 reports on a difference-in-differences analysis comparing performance on open-ended and multiple-choice (MC) items at endline. For each subject, we construct an outcome equal to the difference in the share of correct responses on open-ended items relative to MC items, and estimate a DiD specification contrasting Pioneer and control students. A positive coefficient therefore, indicates that Pioneer students perform relatively better on open-ended items than on MC items, compared to the control group. We interpret these estimates as capturing systematic differences across item formats. Such differences may arise from multiple mechanisms, including differential skill expression across formats or greater subjectivity in scoring open-ended items, particularly in language subjects. While these patterns could also be consistent with some degree of manipulation, the estimated magnitudes are small, and the graphical evidence does not suggest large or pervasive manipulation. Overall, there is no evidence of systematic manipulation, but the language subjects show more room for scorer discretion, which is typical for open-ended linguistic items.

Figure B2: *Manipulation checks comparing open-ended and multiple-choice items across subjects at endline.*



(a) *Arabic*

(b) *French*

(c) *Math*

(d) *Science*

Notes: Each panel plots school-level mean scores on open-ended items against mean scores on multiple-choice items at endline. The dashed 45-degree line indicates equality between the two formats.

B.3.2 Quality control

We implemented a Quality Control (QC) protocol during the endline data collection. This protocol ensures that observed differences in student learning outcomes are attributable to the intervention rather than heterogeneous testing conditions or systematic measurement error between treatment arms.

During endline data collection, we implemented a structured protocol to monitor test administration across schools. Each observer (one per region) was assigned three schools: two Pioneer schools and one Non-pioneer school, and followed a pre-specified sampling plan indicating the subject and grade of the tests to be observed. Observers visited schools during test administration and completed a standardized observation survey covering multiple aspects of implementation quality.

Table B2: *Difference in open-ended vs. multiple-choice scores by treatment status*

	Arabic	French	Science
Pioneer	-0.009**	-0.017***	0.075***
	(0.004)	(0.006)	(0.008)
Observations	4492	4467	4764
R-squared	0.771	0.147	0.121

Notes: Each column reports the estimated difference between Pioneer and non-Pioneer schools in the within-student gap between open-ended and multiple-choice (MC) test items, conditional on grade and triplet fixed effects. The dependent variable, *gap*, measures the difference in the average percentage of correct responses on open-ended and MC items within the same subject and student. A positive coefficient indicates that Pioneer students performed relatively better on open-ended items than on MC items compared to their control-group counterparts. We do not report the coefficient on math as we do not have open-ended math items at baseline. Standard errors are shown in parentheses. Significance levels are denoted as follows: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

We constructed a set of QC indices that capture different dimensions of test administration following a three-stage procedure:

First, individual survey items are recoded so that higher values consistently reflect better data collection quality. Tables B4 and B5 list the individual survey items included in each QC index. For Likert-scale items, we define binary indicators equal to one when responses indicate good or very good quality, and zero otherwise.

Second, subdomain indices are then constructed as simple averages of the relevant recoded items, representing the percentage of quality benchmarks met within that specific area. Specifically, we define subdomains capturing (i) verification of participants, (ii) verification of materials, (iii) logistics and test environment, (iv) examiner protocol compliance (separately for written and oral assessments), (v) training compliance at the enumerator level, (vi) psychosocial and fatigue conditions, and (vii) data reliability.

Finally, these subdomain indices are aggregated into three broad QC dimensions. *Fidelity* combines verification of participants, verification of materials, and training compliance. *Testing Conditions* combines logistics and test environment, psychosocial and fatigue measures, and data reliability. *Protocol Adherence* captures examiner compliance with the testing protocol, using test-typespecific measures for written and oral assessments.

Table B3 reports balance checks for these QC dimensions and their underlying subdomains, comparing Pioneer and control schools. The table reports control and treatment group means and standard deviations, as well as regression-adjusted differences between Pioneer and non-Pioneer schools. Overall, we find no evidence of systematic differences in test administration quality across treatment and control schools, suggesting that test administration quality was comparable across groups. At the subdomain level, we observe a statistically significant difference for Logistics and the test environment. This difference reflects slightly weaker logistical conditions in treatment schools, driven by the availability of waiting rooms for students. Importantly, this difference is limited in magnitude and does not extend to other dimensions of test administration quality. In particular, we find no significant differences between Pioneer and control schools in all the other items.

In addition, we conducted back-checks on a randomly selected subset of tests to assess data reliability. Back-checks consisted of independently re-entering test information and comparing the re-entered data to the original records collected in the field. We construct a binary discrepancy indicator equal to one if any mismatch is detected between the original and back-checked entries. We test for differences in discrepancy rates between Pioneer and control schools by regressing the discrepancy indicator on a Pioneer indicator, clustering standard errors at the triplet level. We find no statistically significant differences in back-check discrepancies across treatment status. Discrepancy rates are low overall, and there is no significant difference across Pioneer Schools and comparison schools.

Results are similar when the analysis is conducted separately by subject, with no systematic pattern of higher discrepancy rates in Pioneer schools. If anything, discrepancies tend to be weakly lower in treatment schools across most subjects. These findings indicate comparable levels of data accuracy across treatment and control schools and suggest that differential measurement error is unlikely to drive the main results.

Table B3: *Balance checks on quality control dimensions*

	Control		Treatment		All
	Mean (SD)	N	Mean (SD)	N	Diff. means (SE)
Fidelity	0.923 (0.150)	53	0.980 (0.062)	118	0.057 (0.034)
Verification of participants	0.893 (0.194)	53	0.966 (0.101)	118	0.073 (0.047)
Verification of materials	0.972 (0.117)	53	0.992 (0.065)	118	0.020 (0.019)
Training compliance	0.906 (0.295)	53	0.983 (0.130)	118	0.077 (0.048)
Protocol Adherence	0.847 (0.120)	53	0.865 (0.088)	118	0.019 (0.021)
Protocol compliance (written)	0.765 (0.157)	17	0.818 (0.141)	22	0.053 (0.069)
Protocol compliance (oral)	0.885 (0.073)	36	0.876 (0.068)	96	-0.009 (0.024)
Testing Conditions	0.912 (0.160)	53	0.902 (0.144)	118	-0.009 (0.035)
Logistics & test environment	0.950 (0.084)	53	0.887 (0.149)	118	-0.063** (0.028)
Psychosocial / fatigue	0.898 (0.182)	53	0.947 (0.129)	118	0.049 (0.028)
Data reliability	0.887 (0.320)	53	0.873 (0.335)	118	-0.014 (0.061)

Notes. This table reports balance checks for quality control (QC) measures constructed from test-level observation data. Rows are organized by three broad dimensions: Fidelity, Testing Conditions, and Protocol Adherence. Each dimension shown is followed by its underlying subdomain measures. For each row, the table reports the control-group mean and standard deviation, the treatment-group mean and standard deviation, and the difference in means between Pioneer (treatment) and control schools. All indices and subdomain measures are coded so that higher values correspond to better quality. Differences in means are estimated from OLS regressions of each outcome on an indicator for Pioneer schools. Standard errors, shown in parentheses, are clustered at the observer level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B4: *Quality control survey items by subdomain*

	Control	**Diff. (SE)**
Data reliability		
Overall data quality of the test session	0.887 (0.320)	-0.014 (0.061)
Logistics and test environment		
Assessment kits available and organized	0.925 (0.267)	-0.026 (0.025)
Booklets complete and in good condition	0.981 (0.137)	0.019 (0.019)
Schedule followed	1.000 (0.000)	-0.025 (0.025)
Quiet and dedicated testing room	1.000 (0.000)	-0.093 (0.061)
Waiting room available	0.943 (0.233)	-0.335** (0.118)
Principal supported test administration	0.849 (0.361)	0.075 (0.072)
Psychosocial and fatigue measures		
Examiner fatigue	0.868 (0.342)	0.064 (0.060)
Student fatigue	0.868 (0.342)	0.047 (0.054)
Student comfort	0.925 (0.267)	0.042 (0.029)
Student did well	0.887 (0.320)	0.054 (0.053)
Good test session overall	0.943 (0.233)	0.040 (0.029)
Training compliance		
Attended full enumerator training	0.906 (0.295)	0.077 (0.048)
Verification of materials		
Required tools available	0.981 (0.137)	0.019 (0.019)
Electronics/calculators forbidden	0.962 (0.192)	0.021 (0.033)
Verification of participants		
Received official student list	0.943 (0.233)	0.057 (0.058)
Verified student identity	1.000 (0.000)	-0.008 (0.008)
Absences documented	0.736 (0.445)	0.171 (0.099)

Notes. This table summarizes the conceptual content of the quality control survey items used to construct each subdomain. The wording shown reflects the main idea of each item and is not the exact phrasing or official translation of the original questionnaire.

Table B5: *Quality control survey items by subdomain (cont.)*

	Control	**Diff. (SE)**
Protocol Adherence: common		
Purpose explained clearly	0.925 (0.267)	0.033 (0.034)
Trust-building and encouragement	0.906 (0.295)	0.060 (0.048)
Time adherence	1.000 (0.000)	-0.017 (0.012)
Students did not leave the room	0.075 (0.267)	-0.025 (0.067)
No unauthorized person entered	0.038 (0.192)	-0.012 (0.042)
Protocol Adherence: written		
Correct booklet assigned	1.000 (0.000)	0.000 (0.000)
Instructions explained	1.000 (0.000)	-0.091 (0.085)
Time indicated	1.000 (0.000)	0.000 (0.000)
Examiner monitored progress	0.647 (0.493)	0.262 (0.213)
Answered procedural questions	0.941 (0.243)	-0.032 (0.117)
No extra guidance on content	0.941 (0.243)	-0.032 (0.117)
Encouraged perseverance	0.588 (0.507)	0.321 (0.203)
Examiner alone with students	1.000 (0.000)	-0.045 (0.043)
Grading followed guide	0.765 (0.437)	0.008 (0.240)
Protocol Adherence: oral		
Examiner conducted oral assessment	1.000 (0.000)	-0.052 (0.050)
French mastery	0.944 (0.236)	-0.144 (0.109)
Introduced self and built rapport	0.917 (0.280)	0.052 (0.066)
Explained structure and duration	0.889 (0.319)	-0.056 (0.128)
Recorded answers during oral test	1.000 (0.000)	-0.021 (0.016)
Read text clearly / used recording	1.000 (0.000)	-0.010 (0.010)
Provided options clearly	1.000 (0.000)	-0.021 (0.021)
No answers to content questions	1.000 (0.000)	0.000 (0.000)
Discreet grading during oral test	0.972 (0.167)	-0.014 (0.041)
Grading followed guide	0.944 (0.232)	0.045 (0.057)
Explained originality requirement	0.917 (0.280)	-0.021 (0.105)
Counted different ideas correctly	0.917 (0.280)	0.083 (0.058)
Encouraged additional ideas	0.972 (0.167)	-0.045 (0.065)
Provided detachable sheet and pencil	1.000 (0.000)	-0.031 (0.024)
Identified drawing elements correctly	0.917 (0.280)	0.083 (0.061)
Named all drawing elements	0.917 (0.280)	0.073 (0.062)
Mentioned all response choices	1.000 (0.000)	-0.010 (0.010)
Managed student fatigue / pacing	0.972 (0.167)	-0.128 (0.117)
Note-taking did not disturb student	1.000 (0.000)	-0.010 (0.011)

Notes. This table summarizes the conceptual content of the quality control survey items used to construct each subdomain. The wording shown reflects the main idea of each item and is not the exact phrasing or official translation of the original questionnaire.