The Best Buy?

From Global Evidence to Government Practice

Hosam Ibrahim ⓡ

Andreas de Barros ⓡ

Sarah Deschênes ⓡ

Paul Glewwe

July 26, 2026

The Best Buy?
From Global Evidence to Government Practice*

Hosam Ibrahim[†]ⓡ

Andreas de Barros[‡]ⓡ

Sarah Deschênes[§]ⓡ

Paul Glewwe[¶]

July 26, 2026

Can governments translate evidence on "what works" into learning gains in public schools? Morocco combined targeted remediation and structured pedagogy (two approaches ranked as "great buys" by a global evidence panel) and delivered them through the existing public-school workforce in 626 schools. Using a pre-registered matched difference-in-differences design across 276 schools and 22,846 students, we estimate that the program increased scores on an independent assessment after one year by 0.90 standard deviations: 0.52 in Arabic, 1.30 in French, and 0.93 in mathematics. Administrative records also show higher grade promotion (1.7 percentage points), while enrollment, which is nearly universal, remained unchanged.

Keywords: Education; human capital; Morocco.

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

Study pre-registration: Registered with a pre-analysis plan at the OSF Registry.

*We thank the Ministry of National Education, Preschool and Sports in Morocco for supporting the impact evaluation of the Pioneer School Program in primary schools. We thank Morocco's *Instance Nationale d'Évaluation* (INE) for making data available. We are grateful to the Morocco Innovation and Evaluation Lab (MEL) and J-PAL staff—Lina Abdelgheffar, Anubhav Agarwal, Youssef Assarssah, Florencia Devoto, Fatine Guedira, Sara Hassani, Rashi Maheshwari, Najiba Mida, and Andrea Salem—for their excellent field coordination, research assistance, and project leadership. We thank Mathilde Col and Quentin Daviot of EVAL-LAB for their excellent support in the data collection process. We thank Imane Fahli from 1SD Impact Consulting and Sanae El Ouarti from the Tony Blair Institute for their invaluable support and commitment, which were instrumental in ensuring the successful implementation of our research project. We also thank Magdalena Bennett, Lelys Dinarte-Diaz, David McKenzie, Matthew Lenard, and Cindy Rojas A. for their constructive comments. Rebecca Thornton supported the study in its early stages. Data protection was secured in compliance with the relevant local legal provisions and regulations concerning Human Subjects research at the University of California, Irvine, the Paris School of Economics, and the University of Minnesota. Following common journal policies, the authors 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. Co-authorship is shared equally among all four authors. The order of the first three authors was determined by a random draw and is indicated by the symbol ⓡ. The findings, interpretations, and conclusions expressed in this paper are those of the authors and do not necessarily represent the views of the World Bank, its Executive Directors, or the governments they represent.

[†]Post-doctoral Researcher, School of Economics and Marketing, University of Technology Sydney. E-mail: hosamm.ibrahim@gmail.com.

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

[§]Economist, Africa Gender Innovation Lab, The World Bank. E-mail: sdeschenes@worldbank.org.

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

1 Introduction

More than half of children in low- and middle-income countries do not learn to read with comprehension by age 10 (World Bank, 2019*a*). Similarly, poor performance has been found in students' basic mathematics skills (Angrist et al., 2021; de Barros and Ganimian, 2023). The World Bank (2018) describes this state of affairs as a global "learning crisis". While almost all of these countries have close to 100 percent of their children enrolled in primary school, progress on learning has been disappointing, raising the question of what can be done in those countries where students are learning very little.

Recent research has provided useful findings on what works to increase student learning in low- and middle-income countries. Based on a systematic review of the evidence, the Global Education Evidence Advisory Panel (GEEAP) (GEEAP, 2023) identified "great buy" interventions, which are highly cost-effective and are supported by a strong body of evidence (Angrist et al., 2024). Consistently, however, such promising interventions have proven less effective (or even detrimental) once external supports, such as researcher oversight, are removed and responsibilities are transferred from a non-governmental organization to the government (Vivalt, 2020; Allcott, 2015). This holds for the "great buy" recommendations, including scripted lesson plans (Kerwin and Thornton, 2021; Piper and Dubeck, 2024) and programs that target instruction to a child's learning level (Banerjee et al., 2017; Duflo, Kiessel and Lucas, 2024). Thus, it remains an open question whether successful educational interventions can maintain their effectiveness if they are implemented under government ownership at scale.

This study measures the impact of Morocco's Pioneer School Program (PSP). The PSP combines two of the GEEAP's three "great buy" recommendations on how to address the learning crisis in low- and middle-income countries. The program's main components consist of structured pedagogy (including scripted lessons) and targeted remediation (by student learning level, not grade). To support the delivery of these two components, the program trains and certifies teachers in the two approaches (with a one-time payment tied to that certification), equips classrooms with projectors to display its scripted lessons, and allows schools to assign some teachers to specialize in teaching Arabic, French, and mathematics. Program implementation is monitored by the Ministry's regular inspectors and successful schools are recognized by the Ministry with a quality-certification label. The program was launched in 626 of Morocco's public primary schools in September of 2023, and this study evaluates the program's impact on student learning over the program's first year. In secondary analyses, we use administrative records covering the universe of students in the study schools to estimate the program's effects on grade promotion and student dropout.

As with the evaluation of any program, credible estimates of the impact of the Pioneer School Program (PSP) require a valid counterfactual. To this end, our prospective, pre-registered study relies on a difference-in-differences strategy. First, leveraging administrative data on the universe of public primary schools in Morocco, we used machine learning methods to obtain a similar non-PSP primary school for each of the PSP schools in the evaluation sample. Then, to estimate the program's impact, we collected primary assessment data and compared changes in student learning over time across these two sets of schools. The key assumption is that the average (conditional) trend in the PSP schools would have been the same as the average (conditional) trend in the matched non-PSP schools had the PSP schools not participated in the program. Comparisons of trends in exam scores over the nine years before the program was implemented support this "parallel trends" assumption.

The estimated intent-to-treat (ITT) impacts of the PSP program are very large. For ease of interpretation, the outcome variables (test scores) have been normalized to have a standard deviation equal to one. Averaging over all three subjects, the impact of the PSP is 0.90 standard deviations (s.d.) of the distribution of the scores of non-PSP schools at endline. By subject, the program's impacts are 0.52 s.d. for Arabic, 1.30 s.d. for French, and 0.93 s.d. for mathematics. The estimated impacts are broadly similar for female and male students, with modest but statistically significant gains for girls in French and on the overall score. Disaggregating by students' within-grade baseline performance, the impacts on Arabic are higher for the bottom quartile than for the top quartile (0.63 versus 0.41 s.d.), the impacts on mathematics are somewhat lower but still very high for the bottom quartile (0.83 versus 0.97 s.d.), and the impacts on French are similar across quartiles (1.31 versus 1.23 s.d.).

Administrative records for the universe of students in the study schools extend these results to schooling outcomes. The program raised the share of students promoted to the next grade by 1.7 percentage points, while enrollment, which is nearly universal in these schools, did not change.

Our first and most direct contribution is to the evidence on what works to improve student learning in low- and middle-income countries, particularly in government-run education systems. The estimated program impact of 0.90 standard deviations (s.d.) is, to our knowledge, the largest ever recorded for a government-implemented program, distinguishing it from programs typically run by non-governmental organizations. Based on a systematic review by Evans and Yuan (2022), this impact places the Pioneer School Program in the top 1 percent of student learning gains for mathematics and reading interventions in low- and middle-income countries: the Arabic impact exceeds the 90th percentile of similar programs, and the French and mathematics impacts exceed the 99th percentile.

Our second contribution is to the literature on state capacity in education. Scaling educational reforms successfully often requires governments to enhance their ability to manage, adapt, and implement programs effectively (Dasgupta and Kapur, 2020; Ganimian, Muralidharan and Walters, 2024). Many successful educational interventions produced their impacts by relying on the non-governmental provision of additional inputs—such as NGO-run summer camps (Banerjee et al., 2017), external organizations hiring extra para-teachers (Eble et al., 2021), governments "outsourcing" the management of public schools to private providers (Romero, Sandefur and Sandholtz, 2020), or other heavy involvement of a government-external party to offer "packaged" interventions with multiple inputs (Maruyama and Igei, 2024).[1] In contrast, attempts to improve government-provided education using existing public sector resources have often struggled to improve student learning (Muralidharan and Singh, 2020; de Barros et al., 2024). In documenting a successful government-implemented program, we show that Morocco's Pioneer School Program achieved significant impacts using only the regular teachers and inspectors already assigned to these schools, with no additional staff or NGO involvement. This provides rare evidence that governments can implement large-scale reforms that improve the efficiency of the public sector and raise student learning without relying on external actors.

Our third contribution is to the understanding of comprehensive educational interventions that integrate structured pedagogy with targeted remediation. Prior studies typically examine these components in isolation, which may overlook the benefits of combining them. 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, pointing to the advantages of multi-pronged educational programs.[2]

Lastly, our results also add to the literature on the effects of remedial education on students' at-grade skills. Prior remedial interventions have shown significant success in augmenting students' foundational skills but fell short of elevating students to their enrolled grade level (Muralidharan, Singh and Ganimian, 2019). In contrast, we find the program not only enhanced students' foundational skills but also yielded substantial gains in at-grade-level competencies. This challenges concerns that remedial interventions may produce improvements too small to close the large skills gaps observed in many low- and middle-income countries (Kaffenberger and Pritchett, 2021) and provides evidence

[1] For an example of a privately managed, bundled intervention that includes structured pedagogy, see (Gray-Lobe et al., 2022). Other programs with strong impacts were implemented in contexts where no formal schooling existed prior to an intervention (e.g., Fazzio et al., 2021).

[2] For another argument for multi-faceted "big-push" interventions from outside education, see Banerjee et al. (2015).

that a well-designed intervention can address remediation and grade-level learning simultaneously.

2 Setting

2.1 Public Education in Morocco

Public provision of education, as governed by the Ministry of National Education, Preschool and Sports (MENPS or *Ministère de l'Education Nationale, du Préscolaire et des Sports*), is the most common type of primary school education in Morocco. Even though the share of private enrollment has increased over the recent years, as of 2022, public schools still served 83 percent of primary school students in the country. Primary school enrollment numbers are high, at 99 percent net enrollment. Persistence is high as well, and as of 2021, 98 percent of enrolled students persisted to grade 6, which is the last grade of primary school (UIS-UNESCO, 2024).

This positive development contrasts with low persistence rates into secondary school and with low student performance in primary school. Nearly 2.3 million children dropped out of primary and lower-secondary schools between 2008 and 2017 (World Bank, 2019*b*), and only about 70 percent of students transition from primary to secondary education (Mansouri and Moumine, 2017). Moreover, learning poverty is high among Morocco's children. In the 2021 PIRLS reading assessment, despite progress over the years, 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. Similarly, in the 2019 TIMSS math assessment, Morocco ranked among the lowest-performing countries (out of 64), and more than half of the students (57 percent) did not reach the minimum proficiency benchmark.

2.2 The Pioneer School Program

The Pioneer School Program (PSP) is inspired by the Global Education Evidence Advisory Panel (GEEAP, 2023) and combines two of its three "great buy" recommendations on how to address the learning crisis in low- and middle-income countries. The program was launched in 626 of Morocco's public primary schools in September of 2023, and this paper evaluates the program's impact over its first year.[3] The Ministry considers the program its flagship intervention in primary schools, and by the 2025-26 school year it had scaled

[3]These schools had volunteered to participate in the program and were then selected by the Ministry.

the program to 4,626 public primary schools, roughly 54 percent of the country's public primary schools, reaching more than two million students.

The program's first component consists of targeted remediation for students in grades 2 to 6, following the *Teaching at the Right Level* (TaRL) approach.[4] The Ministry implemented this remedial component during a dedicated, two-month period at the beginning of the 2023-24 school year. Thereafter, the program included weekly follow-up activities throughout the school year.

The second program component consists of training teachers on how to implement structured pedagogy in their regular classes in grades 1 to 6. To this end, the program provided teachers with scripted lesson plans and a sequence of readily prepared classes that teachers deliver through slide decks they receive from the Ministry and display on projectors.

The program supports the delivery of these two components in several ways. Teachers are trained and certified in the two approaches, and schools may assign some of them to specialize in teaching Arabic, French, or mathematics. Teachers who complete this training and earn certification in the two approaches receive a one-time payment of MAD 10,000 (approximately USD 1,000), tied to that certification and not to their students' test results. Schools that participate successfully are recognized with a "Pioneer School" label. All of the program's implementers are regular teachers and inspectors already assigned to these schools, with no additional staff assigned and no NGO involvement.

3 Research Methods

Our research design and methods follow a registered pre-analysis plan.[5] The plan prespecified our main analyses; Appendix D records the deviations from, and the additions and corrections to, that plan.

3.1 Data

This study uses three sources of data: administrative data for all of Morocco's public primary schools; baseline and endline assessments of students' performance in Arabic, French, and mathematics; and process monitoring data collected during school visits by

[4]The Ministry refers to its program as TaRL, but TaRL Africa and Pratham (the organizations that promote the *Teaching at the Right Level* program in India and multiple African countries) have not been involved in the program's implementation.

[5]See `https://osf.io/zg5ry/`.

inspectors from Morocco's independent evaluation body, the *Instance Nationale d'Évaluation* (INE).

Morocco's Ministry of National Education, Preschool and Sports granted the authors access to retrospective administrative data on all 21,077 public primary schools in Morocco from 2015 to 2022. The data include more than 248 variables measuring student performance and socioeconomic status for 24 million students, as well as many variables on the characteristics of the approximately 104,000 teachers in these schools, and of the schools themselves.

This study's main outcome of interest is student performance on academic assessments in Arabic, French, and mathematics. Baseline assessments were administered in September of 2023, before the program's instruction and remediation began in the classrooms, so they provide a true pre-program baseline; endline assessments were administered in June and July of 2024 (by staff not working in the study schools). For each subject, a two-parameter logistic (2PL) item response theory (IRT) model aggregates students' responses to the test questions (or "items") into continuous estimates of student ability, with overlapping items (or "anchors") placing scores on a common scale across grades and across the baseline and endline rounds.[6] Differential item functioning was investigated, and only items with stable parameters were retained. Each subject's score is standardized to the endline distribution of students in the matched (non-PSP) schools. Appendix B describes the assessments, the item response model, the linking across grades and rounds, and the resulting measurement precision.

The process monitoring data, gathered during announced school visits in April and May of 2024, covers all 626 program schools. We utilize data collected through interviews with the headmasters of all schools, independent observations of each school's infrastructure, and 2,457 classroom observations and teacher interviews (averaging 3.93 observations and interviews per school).

3.2 Sampling

This study's sample of schools was constructed in five steps, using administrative data on the 8,034 government primary schools that formed the matching pool (including the 626 PSP schools). First, using the "post-double-selection" (PDS) methodology (Belloni, Chernozhukov and Hansen, 2014), 17 variables were identified that were predictive of either baseline test scores or participation in the program. Second, using these 17 variables together with the baseline test score (18 variables in total), Mahalanobis nearest-neighbor

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

matching (without replacement) was used to identify a matched (non-PSP) school for each of the 626 PSP schools. Third, post-matching diagnostics identified 496 of the initial 626 matched pairs as "good" matches, yielding a sub-sample of 992 schools from which to sample.[7] Fourth, stratifying by terciles of the grade-6 examination score and by region, 150 of the 496 PSP schools were randomly selected, along with their non-PSP matches. Finally, baseline data collection could not be conducted in 19 of the 300 sampled schools, and five of the remaining 281 schools (three PSP and two non-PSP) thereby lost their matched school; dropping these five yields the study's effective sample of 276 schools (138 PSP and 138 matched non-PSP). Appendix C provides additional technical details on the sampling of schools.

Appendix Table A1 compares the study's program schools with the remaining primary schools in the country, other program schools, and the matched comparison schools. It shows positive selection of schools into the program and into the study. Compared to other schools in the country, the sampled schools are larger, predominantly urban, and serve students who are less poor in terms of their eligibility for social transfer programs, although their average primary school leaving exam scores are slightly lower (Columns 1 to 3).[8] The sample is more representative of other, non-sampled program schools, but it is also more urban and serves students who are less poor (Columns 4 to 6). Yet, as expected from the matching procedure, there are no notable differences between the program schools included in the study and their matched comparison schools (Columns 7 and 8).

The study's unit of analysis is a student. In each school, surveyors were given a "priority" list of five randomly selected students per grade. These lists were constructed before the baseline data collection began, using enrollment records, and they allowed for replacements if students were absent on the day of the baseline assessment. The study's sample of students with baseline test scores includes 22,846 students across the sample of 276 schools (7,706 were tested in Arabic, 7,567 were tested in French, and 7,573 were tested in mathematics). Of these students, 91.0 percent were tracked successfully and took the endline assessments, with no detectable difference in the attrition rate of students between the PSP and non-PSP schools.

[7] Post-matching diagnostics included tests for common support and external validity using the Kernel Probability Density Function (PDF) and Cumulative Distribution Function (CDF) of the grade-6 final examination scores averaged at the school level, and regression-adjusted balance tests of baseline characteristics between treated and matched untreated schools. This process was repeated iteratively, reducing the Mahalanobis distance caliper in increments of 0.1. A final caliper of root 11.5 produced the 496 matched pairs of schools (992 in total), i.e., the "good" matches.

[8] As a proxy for poverty, our school-level administrative data (for the 2021-22 school year) include the share of students qualifying for the conditional cash transfer program "*Tayssir*" (which ended in December of 2023), and a broader measure ("social security") that adds housing subsidies and a free school bag program. We divide a school's number of eligible students by its enrollment, capped at 100 percent (students can be eligible for more than one program).

Table 1 reports four student-level variables at baseline, and attrition rates, separately for the Arabic, French, and mathematics assessments, along with their differences between the PSP and non-PSP schools (column 4). Of the 15 differences (5 variables for 3 assessments), none is significant at the 5 percent level, and only two are significant at the 10 percent level, about what one would expect by random chance; for each assessment, a joint test fails to reject that all five differences are zero (p-values range from 0.24 to 0.85). Appendix Table A2 provides analogous tests for the non-attriting students (those observed at both baseline and endline), with very similar findings.

3.3 Analytical Strategy

We estimate intent-to-treat effects of the PSP intervention by combining difference-in-differences with matching, comparing changes over time in the outcomes of schools assigned to the program with the corresponding changes in their matched comparison schools. For our assessment-based measures of student learning, the following specification is used:

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

Here, Y_{igsjt} is the outcome for student i in grade g in school s in matched pair j in period t, $TREAT_{sj}$ indicates the assignment of school s in pair j to the intervention, and $POST_t$ is a dummy variable equal to 1 if the outcome is measured at endline and 0 if measured at baseline.

The γ_{igsj} term is a student fixed effect, which absorbs all student characteristics, observed or unobserved, that do not change over time; since nearly all students stay in the same school between baseline and endline, it largely absorbs school fixed effects as well. Because the impacts of some characteristics may nonetheless change over time, the specification adds a vector of observed baseline student and school characteristics, X_{igsj}, interacted with $POST_t$, with coefficients δ: students' gender and eligibility for a social safety net measure, and schools' average test scores, region, average class size, and student-teacher ratio. A LASSO determined in a data-driven way which $X_{igsj} \times POST_t$ interactions to include.[9]

The three-way interaction $\zeta(\eta_j \times \theta_g \times \text{POST}_t)$ represents grade-by-matched pair interactions with the post-period indicator; these fixed effects account for how much, on average, a

[9]More specifically, we first calculate each student's change in test scores between the baseline and endline assessments, residualize this change by subtracting the comparison group's grade-by-matched pair mean, and then use a post-double-selection LASSO on these residuals to select relevant predictors X_{igsj} (Belloni, Chernozhukov and Hansen, 2014).

treatment student's same-grade matched-pair comparison peers learned between baseline and endline.

The coefficient of interest, β, captures the intent-to-treat (ITT) effect of assignment to the program, allowing for the possibility that the program was not fully implemented in some PSP schools (none of the non-PSP 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-pair level. To account for multiple hypothesis testing, we present Westfall and Young (1993) stepdown adjusted p-values over a prespecified hierarchy of research hypotheses; additional, exploratory analyses of heterogeneous treatment effects use the Benjamini and Yekutieli (2001) correction, following Anderson (2008).

The identifying assumption is that, conditional on the controls in equation (1), the average trend in the outcomes in the PSP schools would have been the same as in the matched non-PSP schools had the PSP not been implemented. While this assumption cannot be directly tested, Figure 1 shows that average scores on the end-of-primary exams in the two sets of schools are very similar in the nine school years before the PSP was implemented. Appendix Figure A2 plots the program-comparison difference in these scores year by year, overall and by subject; a joint test does not reject that the difference is constant across the pre-program years ($p = 0.64$ overall, and above 0.36 in each subject), consistent with parallel trends.

Equation (1) differences each student's learning between the baseline and endline rounds. Our secondary analyses of administrative outcomes (the enrollment and grade-progression measures in Table 4)[10] cannot use that design: drawn from the Ministry's records for the universe of students in the study schools, these outcomes form repeated cross-sections spanning four school years (2020-21 through 2023-24), with no within-student change between rounds to difference out. We therefore estimate

$$\begin{aligned} Y_{igsjt} = \beta(\text{TREAT}_{sj} \times \text{POST}_t) + \alpha\,\text{TREAT}_{sj} \\ + \delta(X_{igsj} \times \text{POST}_t) + \lambda_{jgt} + \epsilon_{igsjt} \end{aligned} \tag{2}$$

where t now indexes the four school years, POST_t equals one for the program year (2023-24) and zero for the other years, and λ_{jgt} is a matched-pair-by-grade-by-year fixed effect that lets each matched-pair-by-grade cell follow its own trajectory across years, so that the

[10] We observe grade-6 national examination scores for the study schools and use them to gauge pre-program trends (Figure 1) and in a placebo test (Appendix Table A5), but not as a program outcome: in the 2023-24 program year, the Ministry administered a different examination in the Pioneer Schools than in the comparison schools, so post-program examination scores are not comparable across the two groups. In the pre-program years, the examinations were identical for both groups.

three pre-program years identify the counterfactual trend. Without a student fixed effect to absorb it, program assignment enters directly through α, and the vector X_{igsj} now collects only the student's gender, age, and an indicator for prior grade repetition, each interacted with POST_t. As in equation (1), the coefficient of interest β is the intent-to-treat effect, standard errors are clustered at the matched-pair level, and we adjust for multiple hypotheses using the Westfall-Young stepdown procedure. As a placebo check on this design, we also re-estimate equation (2) with the program assigned one school year earlier and the true program year dropped; as shown below, the placebo effects are indistinguishable from zero.

3.4 Compliance

All PSP schools implemented the program. However, during the targeted remediation period at the start of the school year, teacher strikes created substantial variation in schools' ability to carry out the *Teaching at the Right Level* (TaRL) component (see Appendix Figure A1). On average, headteachers reported 32.7 days of TaRL implementation. Although school visits were conducted by Morocco's independent evaluation body, they were announced in advance, so we interpret the classroom observation data with caution. Instead, we focus on infrastructure and indicators less directly influenced by teachers, which they may nevertheless have highlighted in explaining implementation challenges. Infrastructure quality was generally high: all schools were able to conduct lessons with a projector during visits (even if some lessons may have been staged). Only 13.3 percent of schools reported damage to a projector, and while 24.4 percent of schools experienced electricity outages, just 1.4 percent of headmasters described these as occurring more than "occasionally." Lastly, when asked about their comfort with the presentation materials, only 12.1 percent of teachers rated the materials as merely "somewhat" easy to use or more difficult; the vast majority found them clear and straightforward.

4 Results

4.1 Student Learning

The estimated intent-to-treat (ITT) impacts of the PSP program on student learning are given in Table 2. Panel A shows the study's main results, which are aggregated for all students. Since the outcome variables have been normalized to have a standard deviation equal to one, the estimates in Table 2 measure impacts in terms of the standard deviation of the distribution of the endline scores of non-PSP school students.

The results in Panel A show very large impacts. Averaging over all three subjects, the impact of the PSP is 0.90 standard deviations (s.d.), and the subject-specific impacts range from 0.52 s.d. for Arabic to 1.30 s.d. for French, with 0.93 s.d. for mathematics. Compared to the distribution of impacts that Evans and Yuan (2022, Table 2) report for 96 randomized evaluations conducted in low- and middle-income countries, the overall impact exceeds their 99th percentile for the average of reading and mathematics (0.88 s.d.), the Arabic impact is slightly above their 90th percentile for reading (0.50 s.d.), and the French and mathematics impacts are far above their 99th percentiles for reading (0.90 s.d.) and mathematics (0.80 s.d.). These estimates are robust to omitting the data-driven controls and to a raw matched-pair difference in students' baseline-to-endline score changes (Appendix Table A3).

4.2 The Nature of the Learning Gains

The learning gains are broad, and they reach grade-level competencies. Table 3 reports program effects on reading fluency and on the content, cognitive, and curricular-grade-level subdomains to which the assessment items are mapped. The gains are evident even on a concrete skill measured in natural units, independent of the item response theory (IRT) model used to score the main assessments: timed oral reading rose by 7.2 correct words per minute in Arabic and 8.4 in French (Panels A and B), counted directly from students' reading rather than estimated by that model. The gains are also not an artifact of how items were scored. The enumerators who administered the one-on-one assessments were not blind to a school's program status, but restricting the outcome to mechanically-scored items, which admit a single correct answer and leave no room for scorer discretion, leaves the effects large and precisely estimated: 0.34 s.d. in Arabic and 0.86 s.d. in French, with mathematics scored mechanically throughout at 0.93 s.d. (Appendix Table A4).

The gains also reach at-grade material, not only the below-grade foundations that remediation typically targets. Among grade-6 students, assessed on the end-of-primary curriculum, the program raised at-grade scores by 0.51 s.d. in Arabic and 1.08 s.d. in mathematics, with below-grade content improving as well. This sets the program apart from remedial interventions that raise foundational skills without moving students toward their enrolled grade level (Muralidharan, Singh and Ganimian, 2019). The mathematics gains are likewise broad across content: rather than concentrating in the calculation and numeracy domains that remedial programs often emphasize (de Barros and Lubozha, 2026), they are, if anything, slightly larger in geometry, measures, and data (0.97 versus 0.87 s.d.),

and they are as large for items demanding application and reasoning as for those relying on procedural knowledge.[11]

4.3 Heterogeneity

The pre-analysis plan also included secondary analyses by gender and by students' within-grade baseline performance. The impacts disaggregated by gender in Panel B of Table 2 show broadly similar impacts for male and female students, with a modest but significant female advantage in French and on the overall index (both surviving the stepdown adjustment) and no significant difference in Arabic or mathematics. Disaggregating by baseline student performance in Panel C, the impacts on Arabic are higher for students in the bottom quartile (0.63 s.d.) than for students in the top quartile (0.41 s.d.), while the impacts on mathematics are lower (but still very high) for students in the bottom quartile (0.83 s.d.) than for students in the top quartile (0.97 s.d.). For French, the impacts are similar. Overall, the program had very strong effects across gender and across the distribution of baseline student performance.

We also explored heterogeneity along a broader, pre-specified set of student and school characteristics using a causal forest (Athey and Imbens, 2016; Wager and Athey, 2018), following the difference-in-differences implementation of Gavrilova, Langørgen and Zoutman (2025). The results are consistent with the secondary analyses above and add little to them. The program's effects are broadly similar across the characteristics we examine; of the fifteen dimensions tested, only students' baseline performance shows a difference in effect that survives a false-discovery-rate adjustment, with initially lower-performing students gaining somewhat more. Even this gradient is modest, and the least-benefited half of students still gained about 0.80 s.d. Appendix F describes the method and reports the full results.

4.4 Student Progression and Dropout

In Morocco, dropout from public primary school is concentrated at the end of grade 6, when students (fail to) transition from primary to lower-secondary school (Gazeaud and Ricard, 2024). To measure the program's effects on students' progression through school, we apply equation (2) to the Ministry's enrollment records for the universe of students enrolled in the 276 study schools (these administrative outcomes, obtained after the pre-analysis plan was registered, are recorded as additions in Appendix D). Columns

[11] For French, all test questions for students in grades 3, 5, and 6 fall below their enrolled grade level, so the at-grade versus below-grade comparison is not performed for French.

1 to 3 of Table 4 report the program's intent-to-treat effects on three margins: whether a student is enrolled in school, in any grade and including lower-secondary school, in the *following* school year; whether the student is promoted to the next grade at the end of the school year (with students who leave school during the year counted as not promoted); and whether the student remains enrolled at the end of the school year.

Enrollment in these schools is nearly universal: in the comparison group, 96.5 percent of students re-enroll in the following school year, and 97.9 percent remain enrolled through the end of the year. Against these high baselines, the program left both enrollment margins unchanged. The effect on re-enrollment in the following school year, our primary dropout outcome, is 0.10 percentage points (column 1); the confidence interval rules out gains larger than 0.6 percentage points, or about one-sixth of the comparison group's remaining dropout margin. Within-year enrollment is similarly unaffected (column 3). In contrast, the program raised the share of students promoted to the next grade by 1.7 percentage points (column 2), off a comparison-group promotion rate of 92.0 percent; the estimate corresponds to roughly one-fifth of the comparison group's non-promotion rate and survives the Westfall-Young adjustment (adjusted $p < 0.01$).

The promotion gains are broad-based. They are similar for girls and boys (1.6 and 1.8 percentage points, respectively) and for students with and without a history of grade repetition (1.9 and 1.7 percentage points); neither difference in effects is statistically significant. For re-enrollment in the following year, the point estimates are larger for girls than for boys (0.29 versus -0.06 percentage points) and for students with a repetition history (0.63 versus 0.02), but none of these estimates or their contrasts survives adjustment for multiple testing. Because promotion decisions reflect schools' promotion standards as well as students' skills, we do not interpret the promotion effect in isolation; the assessment results in Table 2 indicate that skills themselves improved.

These estimates are unlikely to reflect pre-existing divergence between the two sets of schools. Re-estimating equation (2) with the program assigned one school year earlier, into the pre-program period, and the true program year dropped yields effects indistinguishable from zero, both for the enrollment outcomes above and for scores on the pre-program grade-6 examination (Appendix Table A5), consistent with the parallel trends in Figure 1 and with the absence of anticipation.

5 Conclusion

Formal education can provide many benefits to individuals, but only if they obtain the skills that education is intended to provide. While almost all children in low- and

middle-income countries complete primary school, in many of those countries, they are not acquiring basic reading and mathematics skills. Research in the last two to three decades has revealed which education policies are promising, and many of those countries are now adopting interventions that have been shown to be effective.

Morocco's "Pioneer School Program" (PSP) is one such intervention: it combines two of the Global Education Evidence Advisory Panel's (GEEAP, 2023) three "great buy" recommendations, structured pedagogy and targeted remediation, in the country's public primary schools. While promising, similar education interventions are difficult to scale, and impacts in government schools are often muted. Thus, whether the program would indeed improve learning was an open question.

This paper estimates the impact of the PSP program after one academic year, using a strategy that combines matching methods with difference-in-differences estimation. Exam-score trends in the two sets of schools are very similar over the nine years before the program was implemented, supporting the identifying assumption.

Our study finds that, after one year, the program led to very large improvements in student learning: an intent-to-treat (ITT) impact of 0.90 standard deviations (s.d.) averaged over the three subjects, with 0.52 s.d. for Arabic, 1.30 s.d. for French, and 0.93 s.d. for mathematics. Similarly large, positive impacts hold for female students and for students whose learning levels were farther behind when the program launched, and the results are robust to alternative measures of student learning, including for content and cognitive subdomains not explicitly targeted by the program. Ministry records for the universe of students in the study schools extend the analysis to other schooling outcomes: the share of students promoted to the next grade rose by 1.7 percentage points, while enrollment, which is nearly universal in these schools, did not change.

To our knowledge, the overall program impact of 0.9 s.d. is larger than any impact ever estimated for a program implemented by a government (as distinct from programs implemented by non-governmental organizations), and it exceeds the 99th percentile of treatment effects in the Evans and Yuan (2022, Table 2) review of 96 randomized evaluations in low- and middle-income countries. These findings suggest that combining structured pedagogy and targeted remediation can achieve large impacts on student learning, including when a government delivers the combination at scale through its regular teachers and its own training and certification systems, rather than through external organizations.

Table 1: *Sample of Students and Balance Tests*

	Sample size		Balance	
	Comparison	Program	Comparison	Difference
	(1)	(2)	(3)	(4)
Panel A: Arabic				
% Female	3817	3890	49.78	-1.06
			[50.01]	(1.36)
% Ever repeated	3459	3517	12.98	-0.14
			[33.61]	(0.87)
% Qualified for Tayssir	3817	3890	22.37	0.16
			[41.68]	(1.06)
Baseline test score	3817	3890	-0.37	-0.02
			[0.95]	(0.05)
% Attrited	3817	3890	12.42	-2.81
			[32.98]	(2.26)
Joint F-test (p-value)				0.85
Panel B: French				
% Female	3720	3848	50.05	0.02
			[50.01]	(1.32)
% Ever repeated	3361	3483	12.88	0.72
			[33.51]	(0.93)
% Qualified for Tayssir	3720	3848	23.33	-0.93
			[42.30]	(1.02)
Baseline test score	3657	3800	0.19	-0.11*
			[1.02]	(0.06)
% Attrited	3720	3848	10.51	-2.11
			[30.67]	(1.59)
Joint F-test (p-value)				0.29
Panel C: Math				
% Female	3729	3844	47.98	0.38
			[49.97]	(1.43)
% Ever repeated	3399	3536	13.65	-1.43
			[34.34]	(0.90)
% Qualified for Tayssir	3729	3844	21.99	1.31
			[41.42]	(0.96)
Baseline test score	3729	3844	-0.52	-0.06*
			[0.88]	(0.03)
% Attrited	3729	3844	7.51	-1.38
			[26.36]	(1.21)
Joint F-test (p-value)				0.24

Notes. Sample and unit of observation: 22,846 assessed students across 276 government primary schools. This table reports on the study's sample of students observed at baseline. "Baseline test score" refers to a student's Arabic, French, or mathematics test score at baseline. Each subject's score is standardized to the endline distribution of students in the matched (non-PSP) schools. "Program" and "comparison" refer to 138 Pioneer schools and 138 matched comparison schools, respectively. Difference reports on the regression-adjusted difference between Pioneer schools and comparison schools, controlling for matched-pair-by-grade fixed effects. Standard errors are clustered at the matched-pair level. Standard deviations are shown in brackets; standard errors are shown in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 2: *Intent-to-Treat Effects on Student Learning*

	Overall	By subject		
		Arabic	French	Math
	(1)	(2)	(3)	(4)
Panel A: Overall				
All students	0.90***	0.52***	1.30***	0.93***
	(0.03)	(0.05)	(0.05)	(0.04)
		[0.000]	[0.000]	[0.000]
Panel B: By gender				
Female	0.94***	0.51***	1.34***	0.94***
	(0.03)	(0.05)	(0.05)	(0.04)
	[0.000]	[0.000]	[0.000]	[0.000]
Male	0.87***	0.52***	1.23***	0.90***
	(0.03)	(0.05)	(0.06)	(0.04)
	[0.000]	[0.000]	[0.000]	[0.000]
Female vs male	0.07***	0.01	0.13**	0.03
	(0.03)	(0.04)	(0.05)	(0.04)
	[0.002]	[0.868]	[0.014]	[0.800]
Panel C: By baseline performance				
Bottom quartile	0.92***	0.63***	1.31***	0.83***
	(0.05)	(0.07)	(0.09)	(0.07)
	[0.000]	[0.000]	[0.000]	[0.000]
Top quartile	0.88***	0.41***	1.23***	0.97***
	(0.03)	(0.06)	(0.09)	(0.06)
	[0.000]	[0.000]	[0.000]	[0.000]
Bottom vs top	0.01	0.17**	-0.04	-0.16**
	(0.05)	(0.08)	(0.10)	(0.07)
	[0.881]	[0.043]	[0.868]	[0.043]

Notes. Sample and unit of observation: 20,784 assessed students across 276 government primary schools. This table reports on the program's intent-to-treat (ITT) effects among all students (Panel A), by gender (Panel B), and by students' baseline learning level as per the distribution within their enrolled grade level (Panel C). ITT estimates follow equation (1) of the pre-analysis plan. The last row of Panels B and C reports the difference in effects between the two subgroups, estimated as a pooled treatment-by-subgroup interaction; this row is the test of heterogeneity and need not equal the difference of the two subgroup estimates above it. Column (1) stacks the three subsamples and uses their respective test score as the outcome (Arabic, French, or mathematics). Each subject's score is standardized to the endline distribution of students in the matched (non-PSP) schools. Standard errors are clustered at the matched-pair level and shown in parentheses. Westfall-Young stepdown adjusted p-values, which control the family-wise error rate within each family of tests and are computed by cluster bootstrap on the matched pair, are shown in brackets; the overall (stacked) score for all students is the primary index and is reported without adjustment. Appendix E details the families. * significant at 10%; ** significant at 5%; *** significant at 1%, before adjustment for MHT.

Table 3: *Effects on Subdomains of Student Learning*

	ITT effect	
	All grades	Grade 6 only
Panel A: Arabic		
Words read correctly / minute	7.24***	13.85***
	(0.71)	(1.53)
At grade	0.54***	0.51***
	(0.05)	(0.07)
Below grade	0.39***	0.54***
	(0.05)	(0.07)
Panel B: French		
Words read correctly / minute	8.43***	9.50***
	(2.48)	(2.51)
Panel C: Mathematics		
At grade	0.97***	1.08***
	(0.04)	(0.06)
Below grade	0.87***	0.80***
	(0.04)	(0.06)
Calculation and numeracy	0.87***	0.79***
	(0.04)	(0.06)
Geometry, measures, and data	0.97***	1.08***
	(0.04)	(0.06)
Knowing	0.87***	0.77***
	(0.04)	(0.08)
Applying and reasoning	0.83***	0.88***
	(0.04)	(0.05)
Equal weights to content domains	0.96***	0.98***
	(0.04)	(0.06)

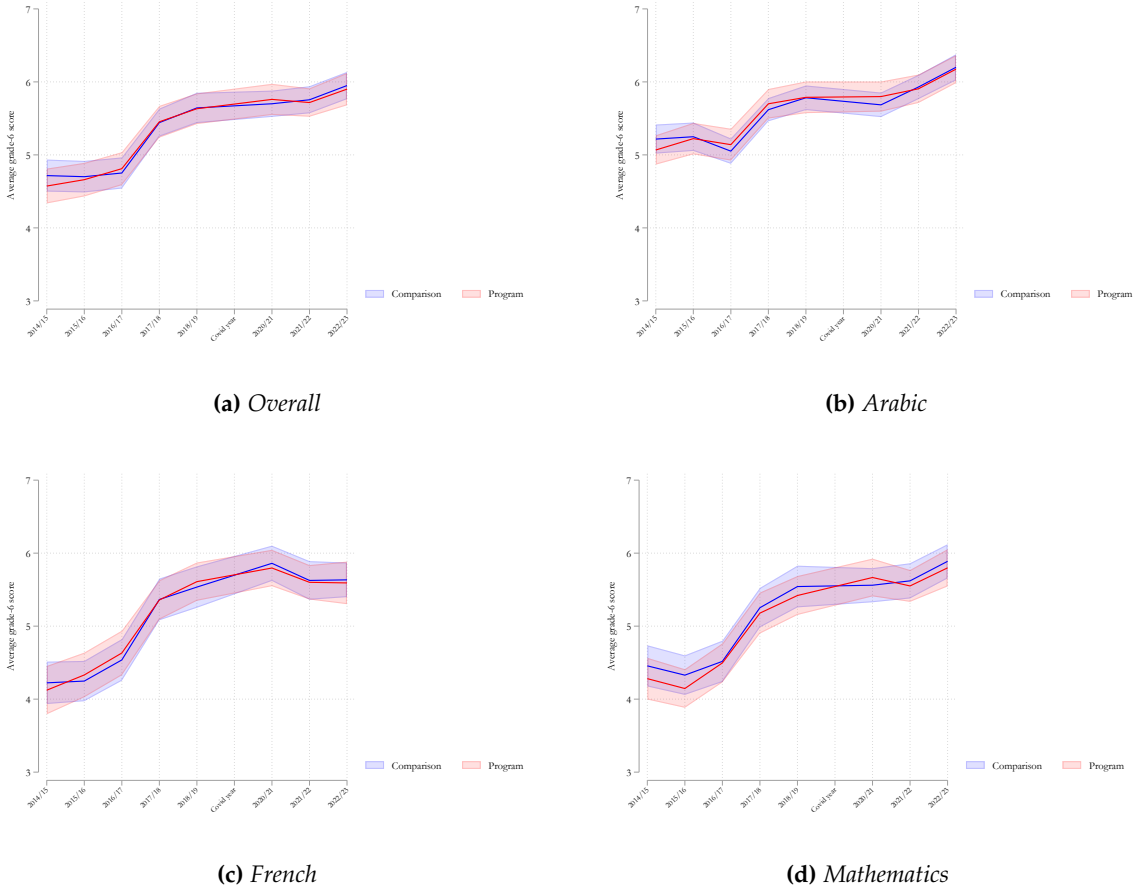
Notes. This table reports on the program's intent-to-treat (ITT) effects by curricular grade level, content subdomains, and cognitive subdomains to which the items are mapped. "At" vs "Below grade" refer to the curricular grade level mapping of test questions. In French, for students in grades 3, 5, and 6, all test questions were below grade level, so we did not perform this additional analysis by curricular grade level. "Equal weights" averages test scores across the two content subdomains (calculation and numeracy vs. geometry, measures, and data domains). ITT estimates follow equation (1). The first column includes all students in grades 1 to 6; the second restricts the same specification to grade-6 students, with cells left blank where an outcome is not assessed at grade 6. Standard errors are clustered at the matched-pair level and shown in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table 4: *Intent-to-Treat Effects on Enrollment and Grade Progression*

	Enrolled next year	Promoted	Enrolled this year
	(1)	(2)	(3)
Panel A: Overall			
All students	0.10	1.66***	-0.03
	(0.24)	(0.37)	(0.20)
		[0.000]	[0.894]
Panel B: By gender			
Female	0.29	1.56***	0.04
	(0.24)	(0.34)	(0.21)
	[0.398]	[0.000]	[0.875]
Male	-0.06	1.75***	-0.09
	(0.27)	(0.50)	(0.22)
	[0.934]	[0.004]	[0.875]
Female vs male	0.34*	-0.17	0.14
	(0.20)	(0.41)	(0.15)
	[0.167]	[0.899]	[0.726]
Panel C: Ever vs never repeated			
Ever repeated	0.63	1.86**	0.46
	(0.44)	(0.82)	(0.29)
	[0.350]	[0.102]	[0.280]
Never repeated	0.02	1.66***	-0.12
	(0.25)	(0.35)	(0.22)
	[0.944]	[0.000]	[0.808]
Ever vs never repeated	0.37	-0.01	0.54*
	(0.44)	(0.71)	(0.31)
	[0.415]	[0.994]	[0.281]
Control mean (2023-24)	96.47	92.01	97.85

Notes. Program intent-to-treat (ITT) effects on three enrollment and grade-progression outcomes for all students in the 276 study schools (Panel A), by gender (Panel B), and by whether the student has repeated a grade by that year (Panel C). Each estimate is the coefficient on program status interacted with a post-(2023-24) indicator in a student-level difference-in-differences over 2020-21 to 2023-24, with matched-pair-by-grade-by-year cell fixed effects and the covariates gender, age, and prior grade repetition interacted with the post indicator. Columns are percentage points: re-enrollment the following year (column 1, the primary dropout outcome and the group's index), grade promotion (column 2), and end-of-year enrollment (column 3). The last row is the 2023-24 comparison-group mean. Standard errors are clustered at the matched pair (parentheses); Panels B and C add a difference-in-effects row (female vs male; ever vs never repeated), estimated as a pooled treatment-by-subgroup interaction; this row is the test of heterogeneity and need not equal the difference of the two subgroup estimates above it. Brackets are Westfall-Young stepdown adjusted p-values, reported unadjusted for the index among all students (Appendix E); these outcomes were added after the pre-analysis plan (Appendix D). * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Figure 1: *Parallel Trends in Grade-6 Examination Scores*



(a) *Overall* **(b)** *Arabic*

(c) *French* **(d)** *Mathematics*

Notes: This figure shows that the program and comparison schools had similar levels and trends in grade-6 examination scores before the program began, overall (panel a) and in each of the three subjects the program targets (panels b–d). Each panel plots, for the 138 program and 138 comparison schools, the yearly average end-of-primary examination score from the 2014-15 through the 2022-23 school year, before the program launched, on a common vertical scale. There is no 2019-20 value because the examination was canceled during the Covid-19 pandemic (the lines connect the 2018-19 and 2020-21 values). Shaded areas are 95 percent confidence intervals, based on standard errors clustered at the matched-pair level.

References

Allcott, Hunt. 2015. "Site Selection Bias in Program Evaluation." *The Quarterly Journal of Economics*, 130(3): 1117–1165.

Anderson, Michael L. 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association*, 103(484): 1481–1495.

Angrist, Noam, David K. Evans, Deon Filmer, Rachel Glennerster, Halsey Rogers, and Shwetlena Sabarwal. 2024. "How to improve education outcomes most efficiently? A review of the evidence using a unified metric." *Journal of Development Economics*, 103382.

Angrist, Noam, Simeon Djankov, Pinelopi K. Goldberg, and Harry A. Patrinos. 2021. "Measuring human capital using global learning data." *Nature*, 592(7854): 403–408. Publisher: Nature Publishing Group.

Athey, Susan, and Guido Imbens. 2016. "Recursive partitioning for heterogeneous causal effects." *Proceedings of the National Academy of Sciences*, 113(27): 7353–7360.

Banerjee, Abhijit, Esther Duflo, Nathanael Goldberg, Dean Karlan, Robert Osei, William Parienté, Jeremy Shapiro, Bram Thuysbaert, and Christopher Udry. 2015. "A multifaceted program causes lasting progress for the very poor: Evidence from six countries." *Science*, 348(6236): 1260799. Publisher: American Association for the Advancement of Science.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton. 2017. "From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application." *Journal of Economic Perspectives*, 31(4): 73–102.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "Inference on Treatment Effects after Selection among High-Dimensional Controls." *The Review of Economic Studies*, 81(2): 608–650.

Benjamini, Yoav, and Daniel Yekutieli. 2001. "The Control of the False Discovery Rate in Multiple Testing under Dependency." *The Annals of Statistics*, 29(4): 1165–1188.

Dasgupta, Aditya, and Devesh Kapur. 2020. "The Political Economy of Bureaucratic Overload: Evidence from Rural Development Officials in India." *American Political Science Review*, 114(4): 1316–1334.

de Barros, Andreas, and Alejandro J. Ganimian. 2023. "The Foundational Math Skills of Indian Children." *Economics of Education Review*, 92: 102336.

de Barros, Andreas, and Theresa Lubozha. 2026. "Targeting Foundational Skills at Scale: Skill Specificity and Transfer." Working Paper 12542, Munich.

de Barros, Andreas, Johanna Fajardo-Gonzalez, Paul Glewwe, and Ashwini Sankar. 2024. "The Limitations of Activity-Based Instruction to Improve the Productivity of Schooling." *The Economic Journal*, 134(659): 959–984.

de Chaisemartin, Clément, and Jaime Ramirez-Cuellar. 2024. "At What Level Should One Cluster Standard Errors in Paired and Small-Strata Experiments?" *American Economic Journal: Applied Economics*, 16(1): 193–212.

Duflo, Annie, Jessica Kiessel, and Adrienne M Lucas. 2024. "Experimental Evidence on Four Policies to Increase Learning at Scale." *The Economic Journal*, ueae003.

Eble, Alex, Chris Frost, Alpha Camara, Baboucarr Bouy, Momodou Bah, Maitri Sivaraman, Pei-Tseng Jenny Hsieh, Chitra Jayanty, Tony Brady, Piotr Gawron, Stijn Vansteelandt, Peter Boone, and Diana Elbourne. 2021. "How much can we remedy very low learning levels in rural parts of low-income countries? Impact and generalizability of a multi-pronged para-teacher intervention from a cluster-randomized trial in the Gambia." *Journal of Development Economics*, 148: 102539.

Evans, David K., and Fei Yuan. 2022. "How Big Are Effect Sizes in International Education Studies?" *Educational Evaluation and Policy Analysis*, 44(3): 532–540.

Fazzio, Ila, Alex Eble, Robin L. Lumsdaine, Peter Boone, Baboucarr Bouy, Pei-Tseng Jenny Hsieh, Chitra Jayanty, Simon Johnson, and Ana Filipa Silva. 2021. "Large learning gains in pockets of extreme poverty: Experimental evidence from Guinea Bissau." *Journal of Public Economics*, 199: 104385.

Ganimian, Alejandro J., Karthik Muralidharan, and Christopher R. Walters. 2024. "Augmenting State Capacity for Child Development: Experimental Evidence from India." *Journal of Political Economy*, 132(5): 1565–1602. Publisher: The University of Chicago Press.

Gavrilova, Evelina, Audun Langørgen, and Floris T. Zoutman. 2025. "Difference-in-Difference Causal Forests With an Application to Payroll Tax Incidence in Norway." *Journal of Applied Econometrics*, 40(7): 727–740.

Gazeaud, Jules, and Claire Ricard. 2024. "Learning effects of conditional cash transfers: The role of class size and composition." *Journal of Development Economics*, 166: 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?" The World Bank, Washington, D.C.

Gray-Lobe, Guthrie, Anthony Keats, Michael Kremer, Isaac Mbiti, and Owen W. Ozier. 2022. "Can Education be Standardized? Evidence from Kenya."

Jacob, Brian, and Jesse Rothstein. 2016. "The Measurement of Student Ability in Modern Assessment Systems." *Journal of Economic Perspectives*, 30(3): 85–108.

Kaffenberger, Michelle, and Lant Pritchett. 2021. "A structured model of the dynamics of student learning in developing countries, with applications to policy." *International Journal of Educational Development*, 82: 102371.

Kerwin, Jason T., and Rebecca L. Thornton. 2021. "Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures." *The Review of Economics and Statistics*, 103(2): 251–264.

Mansouri, Zoulal, and Mohamed El Amine Moumine. 2017. "Primary and Secondary Education in Morocco: From Access to School into Generalization to Dropout." *International Journal of Evaluation and Research in Education (IJERE)*, 6(1): 9.

Maruyama, Takao, and Kengo Igei. 2024. "Community-Wide Support for Primary Students to Improve Foundational Literacy and Numeracy: Empirical Evidence from Madagascar." *Economic Development and Cultural Change*, 72(4): 1963–1992. Publisher: The University of Chicago Press.

Mbiti, Isaac, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Constantine Manda, and Rakesh Rajani. 2019. "Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania." *The Quarterly Journal of Economics*, 134(3): 1627–1673.

Muralidharan, Karthik, Abhijeet Singh, and Alejandro J. Ganimian. 2019. "Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India." *American Economic Review*, 109(4): 1426–1460.

Muralidharan, Karthik, and Abhijeet Singh. 2020. "Improving Public Sector Management at Scale? Experimental Evidence on School Governance India." National Bureau of Economic Research Working Paper 28129, Cambridge, MA.

Piper, Benjamin, and Margaret M. Dubeck. 2024. "Responding to the learning crisis: Structured pedagogy in sub-Saharan Africa." *International Journal of Educational Development*, 109: 103095.

Romero, Mauricio, Justin Sandefur, and Wayne Aaron Sandholtz. 2020. "Outsourcing Education: Experimental Evidence from Liberia." *American Economic Review*, 110(2): 364–400.

UIS-UNESCO. 2024. "UIS statistics."

Vivalt, Eva. 2020. "How Much Can We Generalize From Impact Evaluations?" *Journal of the European Economic Association*, 1–45.

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association*, 113(523): 1228–1242.

Westfall, Peter H., and S. Stanley Young. 1993. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment.* Hoboken, N.J:John Wiley & Sons.

World Bank. 2018. *Learning to Realize Education's Promise. World Development Report 2018.* Washington, D.C.:The World Bank. OCLC: 992735784.

World Bank. 2019*a*. "Ending Learning Poverty: What Will It Take?" The World Bank, Washington, D.C. Publisher: World Bank, Washington, DC.

World Bank. 2019*b*. "Morocco - Education Support Program Project." The World Bank Project Appraisal Document PAD3157, Washington, D.C.

Online Appendix

A Tables and Figures

Table A1: *Sample of Schools, Representativeness, and Balance Tests*

	Representativeness (overall)			Representativeness (within program)			Balance	
	Study	Non-study	Difference	Study PSP	Other PSP	Difference	Comparison	Difference
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Number of teachers	19.52	7.98	11.53***	19.38	17.62	1.76*	19.51	-0.28
	[9.91]	[8.41]	(0.51)	[9.97]	[10.07]	(0.97)	[9.89]	(1.79)
Urban	0.71	0.15	0.56***	0.68	0.58	0.11**	0.72	0.00
	[0.46]	[0.36]	(0.02)	[0.47]	[0.49]	(0.05)	[0.45]	(0.12)
Total enrollment (2021/2022)	534.68	197.29	337.40***	528.65	485.37	43.28	535.42	-43.85
	[314.97]	[256.61]	(15.60)	[305.65]	[320.50]	(30.65)	[325.69]	(75.82)
Female students (percentage, 2021/2022)	48.17	48.05	0.12	48.16	48.17	0.00	48.12	-0.86
	[3.08]	[7.32]	(0.44)	[3.23]	[5.13]	(0.46)	[2.94]	(1.15)
Qualified for Tayssir (percentage, 2021/2022)	17.70	56.22	-38.52***	17.94	25.24	-7.30**	18.12	-1.19
	[28.36]	[29.17]	(1.77)	[27.34]	[31.20]	(2.94)	[29.74]	(6.44)
Qualified for social security (percentage, 2021/2022)	17.81	56.58	-38.77***	18.02	25.45	-7.43**	18.25	-1.19
	[28.53]	[29.34]	(1.78)	[27.36]	[31.51]	(2.96)	[30.04]	(6.44)
Average final grade-6 score (2020/2021)	5.89	5.94	-0.05	5.95	5.81	0.14	5.83	0.14
	[1.00]	[1.15]	(0.07)	[1.07]	[1.02]	(0.10)	[0.92]	(0.22)
Average final grade-6 score (2021/2022)	5.86	5.93	-0.07	5.85	5.93	-0.08	5.88	-0.09
	[0.95]	[1.10]	(0.07)	[0.97]	[0.99]	(0.10)	[0.93]	(0.20)
Average final grade-6 score (2022/2023)	6.02	6.08	-0.05	6.04	5.99	0.05	6.01	0.12
	[0.91]	[1.03]	(0.06)	[0.93]	[0.93]	(0.09)	[0.88]	(0.25)

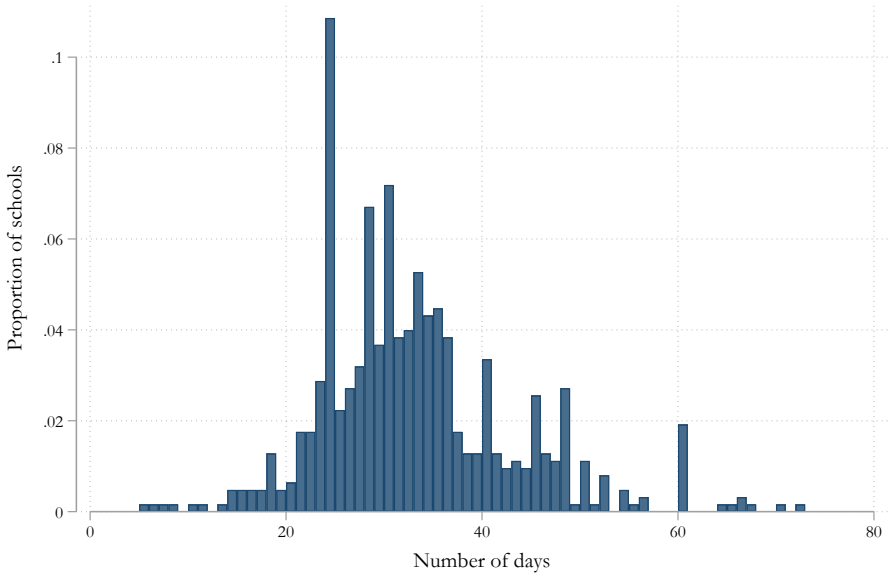
Notes. This table reports on the study's sample of schools. Study refers to the 276 schools included in the study's effective sample. Non-study refers to all other government-primary schools in Morocco. Study (PSP) and Other (PSP) refer to 138 Pioneer schools in the study vs. the remaining 488 Pioneer schools in the country. Comparison refers to the 138 matched comparison schools (vs. the 138 Pioneer schools). Difference reports on the regression-adjusted difference (controlling for matched-pair fixed effects in column 8). The final grade-6 exam score is the school-level average of students' scores on Morocco's national end-of-primary examination, reported for the three post-Covid school years before the program (2020-21, 2021-22, and 2022-23). Standard deviations are shown in brackets; standard errors are shown in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A2: *Sample of Non-attriting Students and Balance Tests*

	Sample size		Balance	
	Comparison	Program	Comparison	Difference
	(1)	(2)	(3)	(4)
Panel A: Arabic				
% Female	3343	3517	49.84	-0.93
			[50.01]	(1.43)
% Ever repeated	3033	3176	12.69	0.11
			[33.30]	(0.92)
% Qualified for Tayssir	3343	3517	22.23	0.71
			[41.58]	(1.18)
Baseline test score	3343	3517	-0.38	-0.01
			[0.95]	(0.06)
Joint F-test (p-value)				0.93
Panel B: French				
% Female	3329	3529	50.23	-0.34
			[50.01]	(1.47)
% Ever repeated	3001	3196	12.86	0.62
			[33.48]	(0.97)
% Qualified for Tayssir	3329	3529	23.19	-0.56
			[42.21]	(1.08)
Baseline test score	3276	3486	0.20	-0.11*
			[1.03]	(0.06)
Joint F-test (p-value)				0.50
Panel C: Math				
% Female	3449	3617	48.13	0.31
			[49.97]	(1.50)
% Ever repeated	3147	3324	13.54	-1.27
			[34.22]	(0.90)
% Qualified for Tayssir	3449	3617	22.09	1.01
			[41.49]	(1.02)
Baseline test score	3449	3617	-0.51	-0.06*
			[0.88]	(0.03)
Joint F-test (p-value)				0.26

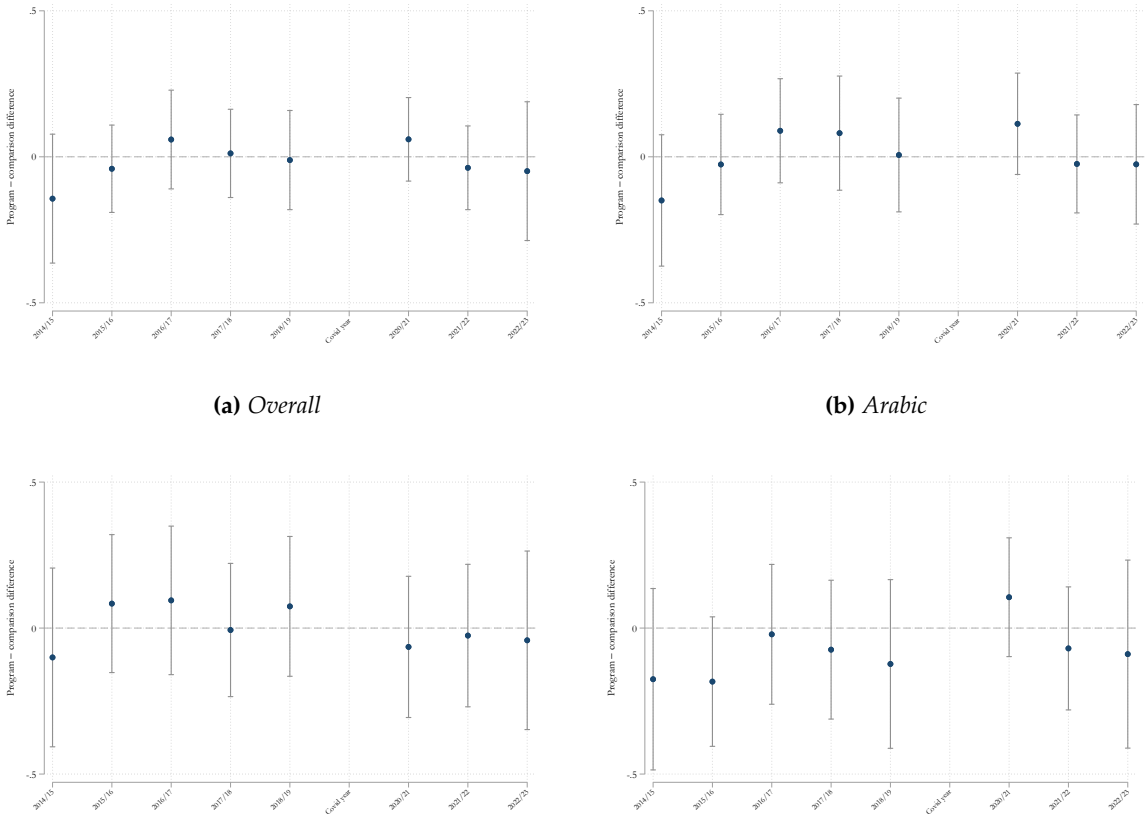
Notes. Sample and unit of observation: 20,784 assessed students across 276 government primary schools. This table reports on the study's sample of non-attriting students (observed at baseline and endline). "Baseline test score" refers to a student's Arabic, French, or mathematics test score at baseline. "Program" and "comparison" refer to 138 Pioneer schools and 138 matched comparison schools, respectively. Difference reports on the regression-adjusted difference between Pioneer schools and comparison schools, controlling for matched-pair-by-grade fixed effects. Standard errors are clustered at the matched-pair level. Standard deviations are shown in brackets; standard errors are shown in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

Figure A1: *School-wise Duration of Teaching at the Right Level Implementation*



Notes: This figure shows substantial heterogeneity in the number of days Pioneer Schools offered the program's *Teaching at the Right Level* component. For the 626 program schools, it presents a histogram with one-day bins summarizing headmaster reports collected during school visits by Morocco's independent evaluation body.

Figure A2: *Pre-Program Trends in Grade-6 Examination Scores: Program-Comparison Differences*



(a) *Overall* **(b)** *Arabic*

(c) *French* **(d)** *Mathematics*

Notes: This figure is the differences counterpart to Figure 1, shown overall (panel a) and for each of the three subjects the program targets (panels b–d). Each panel plots, for every school year from 2014-15 through 2022-23, before the program launched, the program-minus-comparison difference in the average grade-6 national examination score, estimated from a regression of the score on year indicators fully interacted with program status, on a common vertical scale across panels. Vertical bars are 95 percent confidence intervals, based on standard errors clustered at the matched-pair level; the dashed line marks a difference of zero. There is no 2019-20 value because the examination was canceled during the Covid-19 pandemic. A joint Wald test that the differences are equal across years, a test of no differential pre-trend, does not reject in any panel (overall $p = 0.64$, Arabic $p = 0.63$, French $p = 0.86$, mathematics $p = 0.37$).

Table A3: *Robustness of the Learning Effects to the Specification*

	Pre-registered specification	Without selected controls	Raw matched-pair difference
	(1)	(2)	(3)
Overall	0.90***	0.94***	0.94***
	(0.03)	(0.03)	(0.03)
Arabic	0.52***	0.53***	0.54***
	(0.05)	(0.05)	(0.05)
French	1.30***	1.35***	1.36***
	(0.05)	(0.06)	(0.06)
Mathematics	0.93***	0.95***	0.95***
	(0.04)	(0.04)	(0.04)

Notes. Each cell is the program's intent-to-treat effect on the endline learning score, in standard deviations of the endline comparison-group distribution, for the overall (stacked) outcome and by subject. Column 1 is the pre-registered specification whose estimates are reported in Table 2: a student-level difference-in-differences with student fixed effects, matched-pair-by-grade-by-round cell fixed effects, and post-double-selection-LASSO-selected covariates interacted with the endline indicator. Column 2 drops the selected covariates. Column 3 is the raw matched-pair difference, a regression of each student's baseline-to-endline score change on program status with matched-pair fixed effects and no other covariates. Standard errors are clustered at the matched pair (parentheses). * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Table A4: *Effects on Mechanically- and Judgment-Scored Items*

	Mechanically-scored (objective) (1)	Judgment-scored (subjective) (2)
Arabic	0.34***	0.55***
	(0.04)	(0.05)
French	0.86***	1.49***
	(0.04)	(0.05)
Mathematics	0.93***	–
	(0.04)	

Notes. The assessments were administered and scored one-on-one by enumerators who were not blind to a school's program status. To assess whether scorer discretion could account for the effects, each cell reports a separate intent-to-treat estimate using the main specification of Table 2, with the outcome re-scored from the fixed item-response parameters after restricting to mechanically-scored ("objective") items, which have a single correct answer and leave no room for scorer discretion (column 1), or to judgment-scored ("subjective") items such as spoken production, written composition, and open responses (column 2); matched-pair-clustered standard errors are in parentheses. Items are classified from the assessment's endline item map. Effects on mechanically-scored items are large and precisely estimated; mathematics is scored mechanically throughout, so its effect is that reported in Table 2 and it has no judgment-scored counterpart. The larger effects on judgment-scored items are consistent with the program's emphasis on oral and written production, competencies that are inherently judgment-scored. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A5: *Placebo Test for the Administrative Difference-in-Differences Design*

	Enrollment and progression			Overall	Examination, by subject		
	Enrolled next year	Promoted	Enrolled this year	exam score	Arabic	French	Math
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Overall							
All students	0.28	0.44	-0.10	-0.05	-0.04	-0.02	-0.08
	(0.23)	(0.39)	(0.21)	(0.07)	(0.06)	(0.07)	(0.08)
		[0.414]	[0.626]		[0.786]	[0.850]	[0.596]
Panel B: By gender							
Female	0.09	0.43	-0.24	-0.07	-0.05	-0.04	-0.09
	(0.26)	(0.35)	(0.22)	(0.07)	(0.07)	(0.08)	(0.08)
	[0.723]	[0.669]	[0.718]	[0.452]	[0.894]	[0.954]	[0.690]
Male	0.45*	0.39	0.03	-0.04	-0.03	0.01	-0.07
	(0.26)	(0.51)	(0.22)	(0.07)	(0.06)	(0.08)	(0.08)
	[0.240]	[0.861]	[0.898]	[0.723]	[0.979]	[0.997]	[0.834]
Female vs male	-0.32	0.24	-0.21	-0.03	-0.01	-0.05	-0.02
	(0.23)	(0.38)	(0.16)	(0.03)	(0.03)	(0.03)	(0.03)
	[0.290]	[0.750]	[0.545]	[0.412]	[0.958]	[0.381]	[0.911]
Panel C: Ever vs never repeated							
Ever repeated	0.64	0.80	0.18	-0.00	-0.00	-0.02	0.01
	(0.40)	(1.02)	(0.27)	(0.09)	(0.08)	(0.10)	(0.10)
	[0.274]	[0.861]	[0.861]	[0.956]	[0.997]	[0.997]	[0.997]
Never repeated	0.18	0.40	-0.18	-0.07	-0.04	-0.02	-0.11
	(0.24)	(0.33)	(0.23)	(0.06)	(0.06)	(0.07)	(0.07)
	[0.582]	[0.669]	[0.861]	[0.428]	[0.895]	[0.993]	[0.428]
Ever vs never repeated	0.48	-0.20	0.34	0.05	0.01	0.01	0.11*
	(0.43)	(0.88)	(0.30)	(0.05)	(0.06)	(0.06)	(0.06)
	[0.290]	[0.811]	[0.585]	[0.412]	[0.960]	[0.960]	[0.314]
Control mean (2022-23)	96.77	91.68	98.21	0.00	0.00	0.00	0.00

Notes. This table tests the administrative difference-in-differences design (equation 2) for differential pre-trends or anticipation, by moving the treatment one year earlier and dropping the 2023-24 program year, so that the placebo "treatment" falls in 2022-23, a pre-program year, over the window 2020-21 to 2022-23. Columns 1-3 replicate Table 4's enrollment estimator with the treatment shifted. Columns 4-7 apply the same estimator to grade-6 national examination scores; these columns double as a covariate-adjusted, matched-pair-clustered companion to the pre-program trends in Figure 1. Each estimate is the coefficient on program status interacted with the post-(2022-23) indicator in a student-level difference-in-differences, with matched-pair-by-grade-by-year cell fixed effects and the covariates gender, age, and grade repetition interacted with the post indicator. Columns 1-3 are percentage points; columns 4-7 are examination scores, standardized within year against the comparison group. The last row is the 2022-23 comparison-group mean. Standard errors are clustered at the matched pair (parentheses); Panels B and C add a difference-in-effects row (female vs male; ever vs never repeated), estimated as a pooled treatment-by-subgroup interaction. Brackets are Westfall-Young stepdown adjusted p-values, reported unadjusted for the two indices among all students and computed as separate families for the enrollment and examination outcomes (Appendix E), adjusted independently of the main text's confirmatory tests. Under parallel trends and no anticipation, the coefficients should be near zero. * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

B Measurement

The paper's learning outcomes are scores on three assessments, in Arabic, French, and mathematics. This appendix describes the instruments, the item response theory (IRT) model used to score them, how those scores are placed on a common scale across grades and across the baseline and endline testing rounds, and the resulting precision of measurement. IRT models estimate the probability of answering an item correctly as a function of a latent parameter representing the student's ability and of parameters describing the item (Jacob and Rothstein, 2016).

Instruments and Item Development

Each sampled student was assessed in a single subject. Of the students tested at baseline, 7,706 sat the Arabic assessment, 7,567 the French assessment, and 7,573 the mathematics assessment. Students in all six primary grades were assessed, and each grade sat its own test form. In Arabic and mathematics the forms are grade-specific; in French, where the youngest students share more of the curriculum, three forms cover grade 1, grades 2 and 3, and grades 4 through 6. Assessments were administered one-on-one on tablets, with printed stimuli shown to the student, by trained enumerators. A student's assessment lasted about 15 minutes at baseline and about 40 minutes at endline.

The test items were drawn from Morocco's national primary curriculum, and none formed part of the program's own materials: no assessment item appeared in the structured-pedagogy lesson plans and slide decks provided to teachers or in the assessments used to place students for targeted remediation (Teaching at the Right Level). A blueprint maps each item to a content domain, a cognitive domain, and the curricular grade level at which the underlying skill is introduced. In mathematics, the content domains are calculation and numeracy, geometric shapes and measures, and data display and analysis; the cognitive domains distinguish knowledge, application, and reasoning. In Arabic and French, the content domains span oral comprehension and production, decoding and reading fluency, reading comprehension, and written production. This blueprint underlies the subdomain, curricular-grade-level, and cognitive-domain analyses reported in Table 3, each of which rescales a subset of the items on the common metric described below.

The Arabic and French forms also include timed oral-reading passages, from which enumerators recorded the number of words a student read correctly per minute. Because these are counts rather than binary responses, they are summarized directly as reading-fluency measures (Table 3) and are not part of the IRT scaling described here.

Quality Control and Independence

The assessments were administered by trained examiners who did not work in the schools they visited, so that no school's students were assessed by its own staff. Examiners recorded each student's item responses on tablets, and the data-collection software verified each child's identity against the national student register before an assessment could be submitted. The continuous ability scores analyzed in this paper were not generated in the field: examiners recorded students' responses, and the scores were computed centrally by the research team from those responses, using the item response model described below and a fixed set of scoring rules. The study team, through J-PAL's Morocco Evaluation Lab, supervised data quality independently of the schools and of the program, and documented the quality-control procedures described here in a separate report for each round of data collection.

Testing was independently monitored in both rounds. During the baseline and endline assessments, trained inspectors who were independent of the examiners and of the schools observed testing sessions in schools across all twelve regions, covering 63 schools at baseline and 89 at endline. For each session, the inspectors recorded the conditions under which the assessment was administered, including the presence of qualified examiners, the prohibition of calculators and electronic devices, and the classroom environment, and rated the quality of the administration. The conditions of administration were similar in program and comparison schools; calculators and electronic devices, for example, were prohibited in a comparably high share of program and comparison classrooms. The inspectors judged the quality of the assessment data to be high in every subject and in both program and comparison schools.

Item Response Model and Scoring

We score each subject with a two-parameter logistic (2PL) model, which is appropriate for the binary (correct versus incorrect) items that make up the assessments. The 2PL model gives the probability that student j answers item i correctly as

$$P(Y_{ij} = 1 \mid \theta_j) = \frac{1}{1 + \exp[-a_i(\theta_j - b_i)]}, \tag{3}$$

where θ_j is the latent ability of student j, a_i is the discrimination of item i, and b_i is its difficulty. We estimate one model per subject and predict a single continuous ability score for each student using expected a posteriori (EAP) scoring. We then standardize each subject's score by subtracting the mean and dividing by the standard deviation of the score

among comparison-group students at endline, so that treatment effects are expressed in standard deviations of the endline comparison distribution.

Linking across Grades and Testing Rounds

Because students in different grades sit different forms, and because the same students are tested at two points in time, the raw responses do not by themselves place all students on a comparable scale. We use overlapping items, or anchors, to link the forms onto a single scale per subject along two dimensions.

The first dimension is vertical, across grades. Adjacent grades' forms share a block of common items. An item that appears on two grades' forms is scored identically in both, so it acts as an anchor that ties the two grades to the same metric; chaining these overlaps across grades 1 through 6 yields one continuous ability scale per subject that spans all six grades. Each grade's form shares 5 to 8 such items with the adjacent lower grade in mathematics, 7 to 15 in Arabic, and 8 to 12 in French.

The second dimension is temporal, across the baseline and endline rounds. We take the endline data to define the metric: we estimate the item parameters on the endline sample and then hold them fixed while scoring the baseline sample, so that baseline and endline scores share a common origin and unit and changes over time are interpretable. Each endline form shares a substantial block of anchor items with the corresponding baseline form, ranging from 11 to 29 items in Arabic (31 to 82 percent of a form), from 12 to 24 in mathematics (46 to 92 percent), and from 19 to 34 in French (39 to 68 percent).

Linking of either kind is valid only if the anchor items behave the same way in the groups they connect. Before fixing the parameters, we therefore screen each candidate anchor for differential item functioning (DIF), across adjacent grades and across rounds, using both logistic-regression DIF and the Mantel-Haenszel test. Items that function differently across the linked groups are freed, meaning they are allowed group-specific parameters and so do not constrain the linking; only items with stable parameters anchor the scale. This is the sense in which, as noted in the main text, only test questions with stable item parameters are retained for linking.

Precision

Table B1 reports the endline psychometric properties of each assessment: the number of students assessed, the mean item discrimination and difficulty, and the average conditional reliability. The table note records the number of items in each assessment and the average

number administered to a student, since each student sits a single grade form. The assessments measure ability with high reliability, and reliability is high not only on average but across a wide range of ability: Figure B1 plots the standard error of measurement and the conditional reliability implied by the estimated models across the ability scale.[12]

Table B1: *Psychometric Properties of the Assessments (Endline)*

	Students	Mean discrimination	Mean difficulty	Average conditional reliability
	(1)	(2)	(3)	(4)
Arabic	6,860	1.92	0.26	0.92
French	6,858	2.60	-0.28	0.95
Mathematics	7,066	1.98	0.14	0.92

Notes. This table reports the endline psychometric properties of the three student assessments. The unit of observation is a tested student; each student was assessed in a single subject. Properties are estimated from a two-parameter logistic (2PL) item response theory (IRT) model, one per subject, fit to the binary test items. Column (1) gives the number of students assessed at endline; columns (2) and (3) the mean discrimination and difficulty across the model's items; and column (4) the average conditional reliability, the mean across endline students of $\mathrm{Var}(\hat{\theta})/[\mathrm{Var}(\hat{\theta}) + \mathrm{SE}(\hat{\theta}_i)^2]$, where $\hat{\theta}_i$ and $\mathrm{SE}(\hat{\theta}_i)$ are the estimated ability and its predicted standard error for student i. The endline assessments comprise 163 (Arabic), 119 (French), and 145 (mathematics) items, spread across the grade-specific forms and linked by anchor items; each student sits only the items of a single grade form (on average about 32, 40, and 28 items, respectively). Anchor items used to link the endline scale to the baseline scale and across grades are described in the text. Timed reading-fluency items are recorded as word counts and are not part of the 2PL scaling.

References

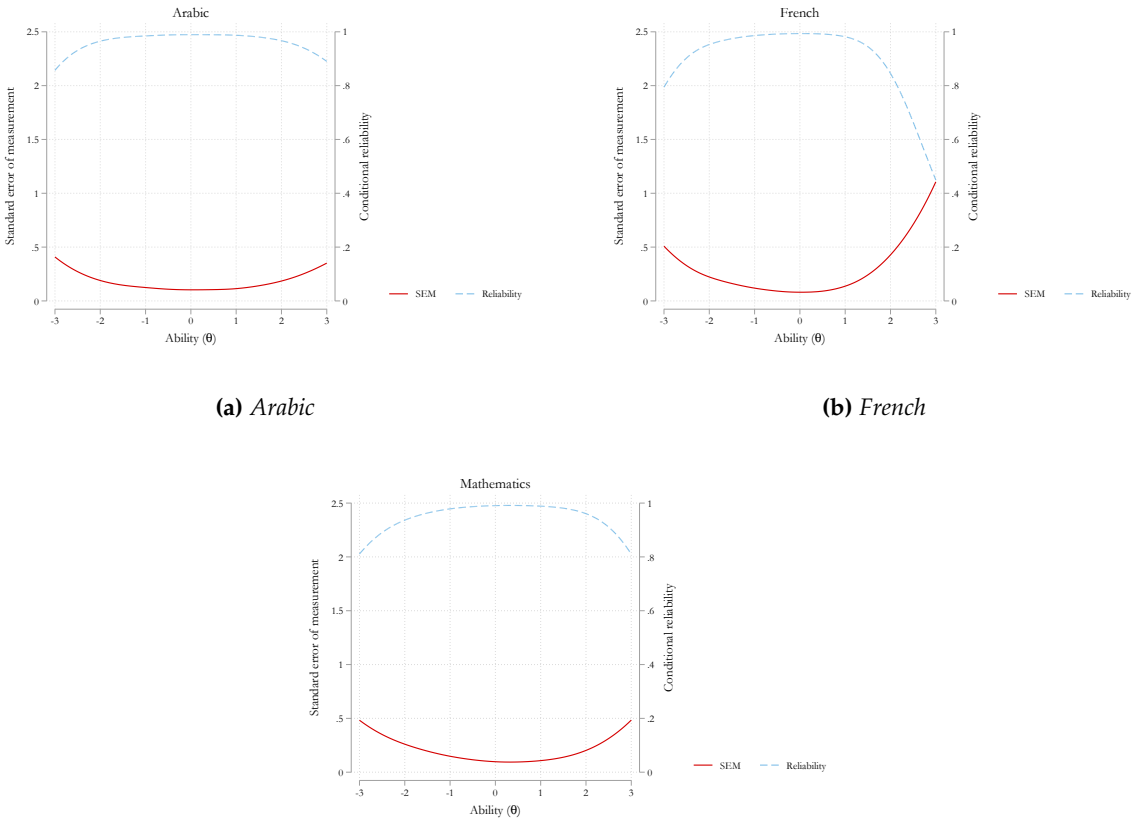
Jacob, Brian, and Jesse Rothstein. 2016. "The Measurement of Student Ability in Modern Assessment Systems." *Journal of Economic Perspectives*, 30(3): 85–108.

[12]We report the average conditional reliability as

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

where N is the endline sample size, $\mathrm{SE}(\hat{\theta}_i)$ is the predicted standard error of the ability estimate for student i, and $\mathrm{Var}(\hat{\theta})$ is the variance of the predicted ability scores.

Figure B1: *Reliability of the Assessments across Ability Levels (Endline)*



(a) *Arabic*

(b) *French*

(c) *Mathematics*

Notes: Each panel reports, for one subject, the precision of the endline ability estimates across the latent ability scale θ. In every panel the solid line is the standard error of measurement, $\text{SEM}(\theta) = 1/\sqrt{I(\theta)}$ (left axis, common 0 to 2.5 scale), and the dashed line is the conditional reliability, $1/[1 + \text{SEM}(\theta)^2]$ (right axis, 0 to 1 scale), both implied by the estimated two-parameter logistic item parameters, where $I(\theta)$ is the test information function. Reliability remains high across a wide range of ability.

C Sampling of Schools

To improve the viability of the conditional parallel trends assumption, we matched schools using the best predictors of treatment status and baseline outcome (Ham and Miratrix, 2024). With access to a large administrative dataset on all primary schools in Morocco, we implemented a machine learning algorithm to identify the most relevant predictors of treatment status and a baseline measure of the outcome of interest (Ham and Miratrix, 2024). The algorithm is post-double selection LASSO (PDSLASSO), used for estimating structural parameters in linear models with many controls (Belloni et al., 2012; Belloni, Chernozhukov and Hansen, 2011, 2014; Belloni et al., 2016).

The PDSLASSO algorithm operates by performing two LASSO regressions. First, it estimates a LASSO regression with the baseline outcome variable as the dependent variable and a set of potential control variables as regressors (125 constructed school-level variables). Second, it performs another LASSO regression with the treatment indicator as the dependent variable and the same set of control variables as regressors. The final set of control variables included in the model is the union of the variables selected in these two steps. This approach ensures that the chosen controls are those most predictive of both the outcome and the treatment assignment (Ahrens, Hansen and Schaffer, 2019).

Using these 125 candidate variables, consisting of school-level characteristics and aggregated student and teacher characteristics for 8,034 primary schools (including the pre-selected program schools), the PDSLASSO identified 17 variables; together with the baseline test score, 18 variables were used for matching.

Next, we implemented an iterative process of a calibrated Mahalanobis distance matching algorithm using these 18 matching variables, to match each of the 626 program schools to an untreated school drawn from the universe of 7,408 untreated schools. After each iteration of matching with different distance calipers, we implemented post-matching diagnostics, including tests for common support, external validity, and balance between the treated and the matched untreated schools. External validity was investigated by comparing the probability density function (PDF) and the cumulative distribution (CDF) for the average baseline measure of the outcome of interest in matched, unmatched, and all schools. Balance in baseline characteristics between the matched treated and untreated schools was evaluated by checking for any statistically significant difference between the paired schools.

The Mahalanobis distance matching was implemented without replacement while restricting the algorithm to select one matched (nearest neighbor) untreated school from the full list of 7,408 untreated schools. Initially, without distance calibration, the matching

algorithm successfully matched 539 treated schools (out of 626) with 539 untreated schools, resulting in a total of 1,078 schools. However, the matched untreated schools from this uncalibrated attempt had significantly different characteristics compared to their matched treated schools, based on the same variables selected by the PDSLASSO. To address this imbalance, we iteratively adjusted the Mahalanobis distance caliper in descending order, starting with the highest possible caliper. We evaluated the baseline balance between schools and examined the PDF and CDF of the baseline measure of the outcome of interest. This process continued until we identified a caliper that produced matched schools with balanced baseline characteristics.

The final caliper used in the matching exercise was root 11.5, producing a list of 496 matched pairs of schools (a total of 992 schools). The population of good matches from the previous step was then used to randomly sample the schools to participate in the baseline and endline data collection for the evaluation of the Pioneer School Program. The evaluation sample was agreed to include 150 treated schools and 150 untreated schools. The evaluation sample was randomly drawn, stratified by terciles of the final grade 6 exam score (the measure for the baseline outcome of interest) and the 12 regions in Morocco, resulting in a total of 36 strata. However, there were 9 strata belonging to the three baseline-outcome terciles in 3 regions with very few schools. As a result, we only stratified the school sampling by region for those 3 regions, with a final total of 30 strata.

Of the 300 sampled schools, baseline data collection could not be conducted in 19 schools (one of which had permanently closed shortly before the study started). Five of the remaining 281 schools (three pioneer schools and two comparison schools) lost their matched schools due to this reduction of the sample to 281 schools. Dropping the five schools without a match resulted in the studys effective sample of 276 schools.

References

Ahrens, Achim, Christian B. Hansen, and Mark E. Schaffer. 2019. "PDSLASSO: Stata Module for Post-Selection and Post-Regularization OLS or IV Estimation and Inference." Statistical Software Components, January.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen. 2012. "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain." *Econometrica*, 80(6): 2369–2429.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2011. "Inference for High-Dimensional Sparse Econometric Models." arXiv.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "High-Dimensional Methods and Inference on Structural and Treatment Effects." *Journal of Economic Perspectives*, 28(2): 29–50.

Belloni, Alexandre, Victor Chernozhukov, Christian Hansen, and Damian Kozbur. 2016. "Inference in High-Dimensional Panel Models With an Application to Gun Control." *Journal of Business & Economic Statistics*, 34(4): 590–605.

Ham, Dae Woong, and Luke Miratrix. 2024. "Benefits and Costs of Matching Prior to a Difference in Differences Analysis When Parallel Trends Does Not Hold." arXiv.

D Deviations from and Corrections to the Pre-analysis Plan

This appendix records where the implemented study departs from the procedures registered in the pre-analysis plan, followed by additions to, and clarifications of, that plan. Most entries are additions that extend the registered analysis using administrative data obtained after registration.

Deviations from the Pre-analysis Plan

- *Measure and sample for student dropout.* The plan registered dropout on the assessed student subsample, together with a subgroup analysis by predicted dropout risk (following the endogenous stratification of Abadie, Chingos and West (2013)). We depart from this in two ways. First, our primary measure of dropout is whether a student re-enrolls the following school year, computed from Ministry enrollment records for the full universe of students in the study schools, rather than end-of-year persistence among the assessed subsample; the assessed-subsample measure is conditioned on assessment participation and records essentially no within-year attrition. Second, we omit the predicted-dropout-risk subgroup, because no baseline covariate predicts dropout in the comparison group (the post-double-selection LASSO selects no predictors); in its place we report a split by whether a student has ever repeated a grade.
- *Standard errors in the school representativeness and balance table.* For the school-level representativeness and balance comparisons in Appendix Table A1, we report homoskedastic standard errors. Because these variables are measured at the school level, we do not cluster; this departs from the plan's table shell, which indicated school-level clustering for the representativeness columns and matched-pair clustering for the balance column.

Additions to, and Clarifications of, the Plan

- *Administrative enrollment and promotion outcomes.* Beyond the registered assessment-based analysis, Table 4 reports intent-to-treat effects on administrative enrollment and promotion outcomes, estimated over the universe of students in the study schools. These administrative data, obtained after the plan was registered, extend the assessment-based learning effects to an independent, administrative outcome. Because these outcomes come from administrative records that form repeated cross-sections, rather than the two-round panel of tested students, they are estimated

with the related specification given in equation (2): matched-pair-by-grade-by-year cell fixed effects (so that each cell follows its own trend), program assignment entering directly in place of a student fixed effect, and the covariates gender, age, and a time-varying grade-repetition indicator (see below), each interacted with the post-program indicator.

- *Grade-repetition indicator.* We measure grade repetition with a time-varying indicator equal to one once a student is observed in the same grade in two consecutive years, computed from the administrative enrollment panel using each student's history up to and including the year of observation. This replaces the lifetime count of repetition years available in the characteristics file, which can include repetitions that occur after, and may be affected by, the program: a student held back at the end of the program year is counted only from the following year, not in the program year itself.

- *Difference-in-effects rows.* In Tables 2 and 4 the subgroup panels include a row reporting the difference in treatment effects between the paired subgroups (girls versus boys, bottom versus top of the baseline distribution, and ever versus never repeated). Each difference is estimated as the coefficient on a treatment-by-moderator interaction in a pooled specification; it is a valid test of heterogeneity but need not equal the difference of the two separately estimated subgroup coefficients reported above it.

- *Parallel-trends figure.* Figure 1 extends the pre-program window from seven to nine school years and is built from student-level administrative grade-6 examination records, rather than school-level averages, with confidence intervals clustered at the matched-pair level; it is shown for the overall examination score and separately for each of the three subjects.

- *School representativeness table.* Appendix Table A1 reports the average grade-6 examination score for three recent school years (2020-21, 2021-22, and 2022-23) in place of the single year in the plan's table shell.

- *Placement of exploratory exhibits.* The exploratory causal-forest analysis of heterogeneous effects (Appendix Table F1) is reported in the appendix rather than the main text.

- *Multiple-hypothesis testing.* We adjust for multiple hypotheses using the pre-registered Westfall-Young stepdown procedure for the confirmatory analyses and the Benjamini-Yekutieli procedure for the exploratory causal forest. Appendix E sets out the graduated family structure through which the registered hierarchy of hypotheses is operationalized. For the causal forest, we additionally summarize the forest's

prioritization rule with the rank-weighted average treatment effect (Yadlowsky et al., 2025), a step not named in the plan.

References

Abadie, Alberto, Matthew M. Chingos, and Martin R. West. 2013. "Endogenous Stratification in Randomized Experiments." National Bureau of Economic Research Working Paper 19742.

Yadlowsky, Steve, Scott Fleming, Nigam Shah, Emma Brunskill, and Stefan Wager. 2025. "Evaluating Treatment Prioritization Rules via Rank-Weighted Average Treatment Effects." *Journal of the American Statistical Association*. arXiv:2111.07966.

E Adjustments for Multiple Hypothesis Testing

We adjust for multiple hypothesis testing following the pre-registered hierarchy of research hypotheses set out in our pre-analysis plan. For the confirmatory analyses of learning and dropout, we report Westfall-Young stepdown adjusted p-values, which control the family-wise error rate within each family of tests (Westfall and Young, 1993). We compute them by resampling, drawing 2,500 bootstrap replications clustered at the matched-pair level, so that the adjustment reflects both the correlation among the tests and the clustering of the design. For the exploratory analysis of heterogeneous effects estimated with a causal forest (Table F1), we instead control the false discovery rate following Benjamini and Yekutieli (2001), as pre-specified for that analysis.

The adjustment is graduated: the more finely we disaggregate an outcome, the larger the family of tests against which a given estimate is adjusted. We organize each outcome into three levels.

(i) *Index.* We first report an aggregate for the outcome. For learning, this is an overall score that stacks a student's Arabic, French, and mathematics tests; for enrollment, it is whether a student re-enrolls in the following school year. The index is the single primary test for the outcome and is reported without adjustment.

(ii) *Components.* We then report the outcome's components. For learning, these are the three subject scores; for enrollment, they are grade promotion and within-year enrollment. The components are adjusted jointly, as one family.

(iii) *Subgroups.* Within each outcome we estimate effects for pre-specified subgroups. The subgroup effects on the index form one family; the subgroup effects on the components form a larger family, and are therefore adjusted more heavily. Adding a subgroup dimension enlarges the family, making the adjustment more conservative.

We also test for heterogeneity directly, through the difference in treatment effects between paired subgroups. For the learning outcomes these are girls versus boys and the bottom versus the top of the baseline distribution; for the enrollment outcomes they are girls versus boys and students who have ever versus never repeated a grade. Because a difference is a distinct hypothesis from either subgroup's effect, the difference tests form their own families, adjusted among themselves at the index and component levels within each outcome.

Our main analysis of learning was pre-registered and is adjusted within the families above. One set of outcomes was added after the pre-analysis plan: administrative analyses of

grade promotion and school enrollment across the universe of study schools. To keep the pre-registered tests separate from this addition, the added outcome is adjusted within its own families, so that it never affects the adjusted p-value of a pre-registered test. For the same reason, the pre-registered subgroup of students at high predicted risk of dropping out, which we could not construct because dropout is not predictable from the observed data, is replaced by a split on whether a student has ever repeated a grade; the resulting subgroup family is the same size as the one we pre-registered.

The placebo test in Appendix Table A5, a falsification check rather than a result, is adjusted separately from the confirmatory hierarchy above, using the same graduated family structure, so that it cannot affect any pre-registered or added p-value.

Our analyses of implementation fidelity and program take-up are descriptive and are not included in these adjustments.

References

Benjamini, Yoav, and Daniel Yekutieli. 2001. "The Control of the False Discovery Rate in Multiple Testing under Dependency." *The Annals of Statistics*, 29(4): 1165–1188.

Westfall, Peter H., and S. Stanley Young. 1993. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment.* Hoboken, N.J:John Wiley & Sons.

F Exploring Heterogeneity with a Causal Forest

This appendix details the causal-forest analysis of treatment-effect heterogeneity summarized in Section 4.3. To better understand who benefited the most from the program, we estimate conditional average treatment effects along a broader, pre-specified set of student and school characteristics using the causal forest of Athey and Imbens (2016) and Wager and Athey (2018).

To accommodate our difference-in-difference setup, following Gavrilova, Langørgen and Zoutman (2025), 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 pair-by-grade fixed effects, the strata of our design. Third, from this regression, we calculated the residuals. Finally, using the "honest" approach (Athey, Tibshirani and Wager, 2019), we trained a causal forest with these residuals and a pre-specified set of covariates (the controls from the main-effect analyses together with the student and school characteristics reported in Table F1), building 50,000 trees, setting the minimum number of treatment and control observations allowed in a leaf to the default value (5), and clustering at the matched-pair level.

Following Carlana, La Ferrara and Pinotti (2022) and Dinarte et al. (2024), we use the out-of-bag predictions to split students at the median of the predicted treatment effect, into a top half (the "Strong" group) and a bottom half (the "Weak" group). To understand what types of students are more likely to be positively affected by the program, we then characterize these groups with a balance test. Table F1 presents the results.

The top row shows that the group-average intent-to-treat effect is large in both groups, at 0.80 s.d. in the "Weak" group and 1.07 s.d. in the "Strong" group, a difference of 0.26 s.d. The rank-weighted average treatment effect test rejects the null of no heterogeneity ($p < 0.001$), so the forest orders students by their predicted benefit in a way that captures systematic variation, even as the program raised learning substantially throughout the distribution. Panel A shows that the variation is concentrated in students' baseline performance: the "Strong" group is composed disproportionately of students in the bottom within-grade quartile, and of students with below-median baseline scores more generally, with conditional average treatment effects of 1.10 and 1.00 s.d. against 0.73 s.d. for the top quartile and 0.87 s.d. for above-median students. Of the fifteen characteristics we examine, only these two baseline measures show a difference in effect that survives the false-discovery-rate adjustment in column (8). The modest over-representation of girls and of upper-grade students in Panel A, and of higher-scoring schools in Panel B, does not survive the adjustment, and the remaining school characteristics are unrelated to the

ranking. Even the baseline gradient is more pronounced in the forest, which projects treatment effects onto the continuous baseline score, than in the discrete within-grade quartile contrasts of Table 2, Panel C, where the bottom- and top-quartile effects are similar. The gradient is therefore suggestive of somewhat larger gains for initially lower-performing students rather than evidence of a large gap, especially because even the least-benefited half gained about 0.80 s.d.

Table F1: *Exploring Conditional Average Treatment Effects on Student Learning*

	Importance	Subgroup indicator	CATE	Standard error	Weak group	Strong group	Difference	q-value (BY)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Student characteristics								
Female	0.003	Yes	0.972	0.033	0.482	0.507	0.025	0.110
		No	0.901	0.032	0.518	0.493	-0.025	
Bottom quartile	0.030		1.103	0.047	0.079	0.426	0.347	<0.001
Middle two quartiles	0.002		0.953	0.032	0.430	0.566	0.136	
Top quartile	0.129		0.733	0.033	0.491	0.009	-0.483	
Grade 1	0.010		1.026	0.062	0.118	0.166	0.048	0.291
Grade 2	0.006		0.840	0.050	0.190	0.134	-0.056	
Grade 3	0.004		0.835	0.045	0.211	0.133	-0.078	
Grade 4	0.007		1.068	0.041	0.141	0.207	0.065	
Grade 5	0.002		0.919	0.044	0.174	0.172	-0.002	
Grade 6	0.001		0.938	0.038	0.166	0.188	0.023	
Baseline score	0.063	Low	1.004	0.038	0.305	0.695	0.390	<0.001
		High	0.868	0.028	0.695	0.305	-0.390	
Grade point average	0.021	Low	0.913	0.032	0.548	0.451	-0.097	0.511
		High	0.959	0.033	0.452	0.549	0.097	
Ever repeated	0.004	Yes	0.903	0.049	0.135	0.123	-0.011	1.000
		No	0.934	0.031	0.865	0.877	0.011	
Ever qualified for Tayssir	0.002	Yes	0.909	0.042	0.231	0.221	-0.010	1.000
		No	0.944	0.029	0.769	0.779	0.010	
Panel B: School characteristics								
Number of teachers	0.088	Low	0.974	0.050	0.401	0.548	0.147	1.000
		High	0.902	0.035	0.599	0.452	-0.147	
Urban	0.002	Yes	0.915	0.032	0.775	0.628	-0.147	1.000
		No	0.986	0.064	0.225	0.372	0.147	
Regional development	0.019	Yes	0.979	0.039	0.577	0.644	0.067	0.511
		No	0.869	0.045	0.423	0.356	-0.067	
Total enrollment (2021/22)	0.142	Low	0.943	0.046	0.428	0.565	0.137	1.000
		High	0.930	0.039	0.572	0.435	-0.137	
Female students (%)	0.102	Low	0.954	0.042	0.494	0.502	0.007	1.000
		High	0.919	0.044	0.506	0.498	-0.007	
Tayssir beneficiaries (%)	0.084	Low	0.935	0.040	0.543	0.457	-0.086	1.000
		High	0.938	0.045	0.457	0.543	0.086	
Social security beneficiaries (%)	0.074	Low	0.937	0.040	0.542	0.457	-0.085	1.000
		High	0.936	0.045	0.458	0.543	0.085	
Average grade-6 exam score	0.207	Low	0.858	0.037	0.631	0.362	-0.269	0.095
		High	1.013	0.044	0.369	0.638	0.269	
Group-average ITT effect					0.804	1.068	0.264	
					(0.030)	(0.038)	(0.049)	
RATE p-value							<0.001	

Notes. Using a causal forest (Wager and Athey, 2018; Athey, Tibshirani and Wager, 2019), this table explores heterogeneity in intent-to-treat effects on student learning, beyond the pre-registered subgroup analyses in Table 2. Panel A reports student-level covariates and Panel B school-level covariates; both come from a single joint forest, so variable importance, the weak/strong split, and the group-average ITT at the bottom are global to that forest. We stack the three subsamples and use the residualized change in test score (Arabic, French, or mathematics) as the outcome; the sample is the attrition-free analysis sample of student-by-subject observations, with missing covariates retained and handled within the forest. The propensity is estimated by the forest (a regression forest on the covariates) rather than fixed at one half: school-level one-of-two assignment does not pin the observation-level treated share to 0.5 because pair enrollments differ. The forest clusters at the matched-pair level. Column (1) reports variable importance, the (depth-weighted) share of times each variable is selected for splitting when building trees; it is a descriptive heuristic that tends to be larger for continuous covariates, which offer more split points. Column (2) indicates the subgroup. Binary characteristics are shown as "Yes"/"No"; the within-grade baseline quartiles and the grade indicators are shown one row per level; and continuous covariates (baseline score, grade point average, and the school-level counts, percentages, and exam score) are dichotomized at their sample median into "Low" (below) and "High" (at or above), so that a subgroup CATE can be reported for each. Baseline quartiles are formed within grade, as in Table 2. Grade point average is missing for a handful of students, who are omitted from that row's subgroup CATEs and share entries. Column (3) reports the conditional average treatment effect (CATE) for each subgroup, from a best linear projection of the forest's out-of-bag CATEs, and column (4) its standard error, clustered at the matched-pair level. In columns (5) and (6), the "weak group" comprises students whose predicted CATE falls below the median of the forest's out-of-bag predictions and the "strong group" those above; entries report the share of a row's subgroup in each group, and column (7) the difference in shares. Column (8) reports the Benjamini-Yekutieli step-up q-value (Benjamini and Yekutieli, 2001) for a test of whether the conditional average treatment effect differs across a characteristic's groups, from a doubly-robust best linear projection of the forest's out-of-bag CATEs on a single indicator, with standard errors clustered at the matched-pair level; the projection slope equals the difference in the two subgroup effects in column (3). The adjustment controls the false discovery rate across the 15 characteristics, as pre-specified for this exploratory analysis, and the q-value is shown once per characteristic. For the two multi-category characteristics it uses a single pre-specified contrast, grades 1–3 versus 4–6 and the top versus bottom baseline quartile (the two middle quartiles omitted), reported on the first row of the block. "Group-average ITT effect" reports the augmented inverse-propensity-weighted intent-to-treat effect within each group (Athey and Wager, 2019), with standard errors in parentheses; the final column reports the difference, with its standard error computed under independence of the two non-overlapping groups. "RATE p-value" reports the two-sided p-value from a sequential cross-validation test of the rank-weighted average treatment effect (AUTOC) (Yadlowsky et al., 2025), using cluster-respecting folds ($K = 5$), which tests whether the forest's treatment-effect ranking captures genuine heterogeneity. 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).

References

Athey, Susan, and Guido Imbens. 2016. "Recursive partitioning for heterogeneous causal effects." *Proceedings of the National Academy of Sciences*, 113(27): 7353–7360.

Athey, Susan, and Stefan Wager. 2019. "Estimating Treatment Effects with Causal Forests: An Application." *Observational Studies*, 5(2): 37–51.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized random forests." *The Annals of Statistics*, 47(2): 1148–1178. Publisher: Institute of Mathematical Statistics.

Benjamini, Yoav, and Daniel Yekutieli. 2001. "The Control of the False Discovery Rate in Multiple Testing under Dependency." *The Annals of Statistics*, 29(4): 1165–1188.

Carlana, Michela, Eliana La Ferrara, and Paolo Pinotti. 2022. "Goals and Gaps: Educational Careers of Immigrant Children." *Econometrica*, 90(1): 1–29.

Dinarte, Lelys, Pablo Egana-delSol, Claudia Martinez A., and Cindy Rojas A. 2024. "When Emotion Regulation Matters: The Efficacy of Socio-Emotional Learning to Address School-Based Violence in Central America."

Gavrilova, Evelina, Audun Langørgen, and Floris T. Zoutman. 2025. "Difference-in-Difference Causal Forests With an Application to Payroll Tax Incidence in Norway." *Journal of Applied Econometrics*, 40(7): 727–740.

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association*, 113(523): 1228–1242.

Yadlowsky, Steve, Scott Fleming, Nigam Shah, Emma Brunskill, and Stefan Wager. 2025. "Evaluating Treatment Prioritization Rules via Rank-Weighted Average Treatment Effects." *Journal of the American Statistical Association*. arXiv:2111.07966.