A collaboration between

The best buy?

Prospective evidence on successful remediation in Morocco's public primary schools

Hosam Ibrahim

Andreas de Barros

Sarah Deschênes

Paul Glewwe

September 28, 2024

The best buy? Prospective evidence on successful remediation in Morocco's public primary schools*

Hosam Ibrahim†

Andreas de Barros‡

Sarah Deschênes§

Paul Glewwe¶

September 28, 2024

To address the learning crisis, targeted remediation and structured pedagogy are often considered "great buys" for education policymakers, but related programs are difficult to scale, and impacts in government schools are often muted. We collaborated with Morocco's Ministry of National Education Preschool and Sports to conduct a prospective evaluation of a large-scale reform effort that combines multiple "great buys". Our results rest on a pre-registered difference-in-differences analysis of primary data collected from 276 public primary schools. After one year, the program improved student learning by 0.90 standard deviations (s.d.) on average (0.52 s.d. impacts in Arabic, 1.30 s.d. impacts in French, and 0.93 s.d. impacts in math). Its average effect exceeds the 99th percentile of treatment effects of other education interventions in low- and middle-income countries. These findings suggest that an integrated intervention can achieve transformative impacts on student learning, at scale.

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 Preschools and Sports in Morocco for supporting the impact evaluation of the Pioneer School Program in primary schools. 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 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. Co-authorship is shared equally among all four authors. The order of the first three authors was determined by a random draw. This paper will undergo peer review. 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.

†PhD student, Department of Applied Economics, University of Minnesota. E-mail: ibrah252@umn.edu

‡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.

¶Distinguished McKnight Professor, Department of Applied Economics, University of Minnesota. E-mail: pglewwe@umn.edu.

Executive summary

Background. During the early 2000s, Morocco had some of the lowest education outcomes compared to other low- and middle-income countries. While Morocco has made remarkable progress in education in recent years, the learning levels of the country's primary school students are still alarming. In the 2021 PIRLS reading assessment, Morocco's fourth-graders ranked second to last (out of 57 countries), and more than half of the students (59 percent) did not reach the minimum proficiency benchmark. 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. This raises the question of what can be done to increase student learning in Morocco.

The Pioneer School Program. This study measures the impact of Morocco's Pioneer School Program (PSP). The PSP is inspired by the Global Education Evidence Advisory Panel (GEEAP, 2023) and combines two of its "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). It complements these program components by assigning some primary school teachers to specialize in teaching Arabic, French, and mathematics and by recognizing schools that successfully participate in the program (with a system of quality certification). The program was launched in 626 of Morocco's public primary schools in September 2023, and this report evaluates the program's impact on student learning over the program's first year. In secondary analyses, we also report on the program's impact on student dropout.

Data. Our study uses three sources of data: 1. Administrative data for all of Morocco's public primary schools; 2. Baseline and endline assessment data that measure students' performance in Arabic, French, and mathematics; and 3. Process monitoring data on program take-up and implementation fidelity. This report's main outcomes of interest are student performance on academic assessments in Arabic, French, and mathematics. Baseline assessments for each of the three subjects were administered in September of 2023, and endline assessments were administered in June and July of 2024. The process monitoring data collected during school visits has yet to be shared with the research team and is not used in this document. We are grateful for the Ministry's commitment to sharing this data with us.

Constructing a comparable sample of schools. As with the evaluation of any program, credible estimates of the impact of the PSP require a valid counterfactual. This study combines difference-in-differences with matching to meet this requirement. First, matching methods were used to obtain a similar non-PSP primary school for each of the 626 schools that were assigned to the PSP in September of 2023. Of the 626 schools that implemented the PSP, 150 (along with their 150 non-PSP matches) were randomly selected for this evaluation. Unfortunately, baseline data could not be collected in 19 schools. In addition, 5 of the remaining 281 schools (3 PSP schools and 2 non-PSP schools) lost their matched schools due to the loss of these 19 schools. Dropping these 5 schools yields the study's sample of 276 schools (138 PSP and 138 matched non-PSP).

Confirming the study's analytical strategy is valid. Changes in outcomes of interest over time were compared over these two sets of schools to estimate the impact of the PSP. The key assumption is that the PSP schools' average (conditional) trend in student performance 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 seven years before the program was implemented support this "parallel trends" assumption (Figure E1).

Confirming that students are similar across groups. The students in the study's 276 schools are the study's unit of analysis. The study has baseline test scores for 22,846 students in these 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 successfully tracked and took the endline assessments, with no difference in the attrition rate of students between the PSP and non-PSP schools. Comparisons of the characteristics of students in the PSP and non-PSP schools indicate that these two sets of students are very similar.

Program impacts. The estimated intent-to-treat (ITT) impacts of the Pioneer School Program indicate extremely large impacts (Table E1). 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 impact is 0.52 s.d. in Arabic, 1.30 s.d. in French, and 0.93 s.d. in mathematics.

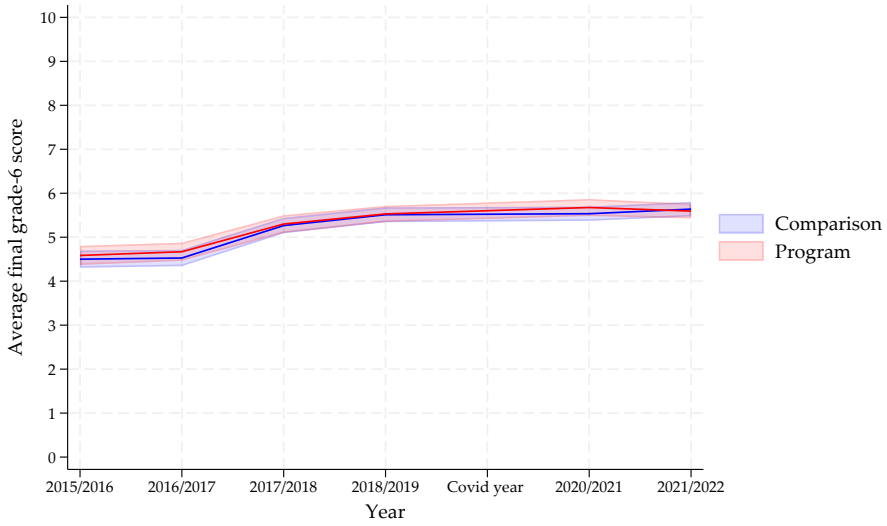
Program impacts among female students and low performers. The estimated impacts are very similar for female and male students, with slightly larger impacts for female

students in French and mathematics and for the overall score. Disaggregating by baseline student performance (within each grade level), the impacts on Arabic are higher for students in the bottom 25 percent of students in a given grade's baseline distribution (0.63 s.d.) than for students in the top 25 percent of that distribution (0.41 s.d.), while the impacts on mathematics are lower (but still extremely high) for students in the bottom 25 percent (0.83 s.d.) than for students in the top 25 percent (0.97 s.d). For French, the impacts are slightly higher for students in the bottom 25 percent of students in that distribution (1.31 s.d.) than for students in the top 25 percent (1.23 s.d.).

Putting the results into perspective. 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). We offer two additional interpretations of the program's impacts. First, the top panel of Figure E2 illustrates that the average student in a PSP school now performs better than about 82 percent of their peers in the comparison schools that did not receive the program (assuming test scores are normally distributed). Second, the bottom panel of Figure E2 shows that, compared to other programs in low- and middle-income countries, the overall impact of 0.90 s.d. is within the top 1 percent of impacts on student learning in mathematics and reading, based on a systematic review of program impacts conducted by Evans and Yuan (2022). In addition (not shown in Figure E2), the 0.52 s.d. impact for Arabic is slightly higher than the 0.50 s.d. impact at the 90th percentile of Evans and Yuan for reading, and the 1.30 s.d. impact for French is much higher than the 0.90 s.d. impact at the 99th percentile of Evans and Yuan for reading. Finally, the impact of 0.93 for mathematics is higher than the 0.80 s.d. impact at the 99th percentile of Evans and Yuan for mathematics.

Outlook. Our next steps consist of adding the data on implementation fidelity and program take-up to our analysis and updating our analysis of student dropout with data that follows the study's sample of students into the next year (including, for grade-6 students, whether they enrolled in secondary school). Also, we will start to disseminate the study's results, including at the upcoming *Forum National de l'Enseignant* in Rabat. In light of the observed impacts, we look forward to seeing the program be scaled up more widely. We are eager to build on this successful collaboration with the Ministry and to engage in similar research studies (with the J-PAL-affiliated Morocco Innovation and Evaluation Lab). If there is interest from the Ministry, this may also include randomized trials to investigate which of the program components drive the success of the Pioneer School Program.

Figure E1: *Program and comparison schools produced similar exam scores before the program started*



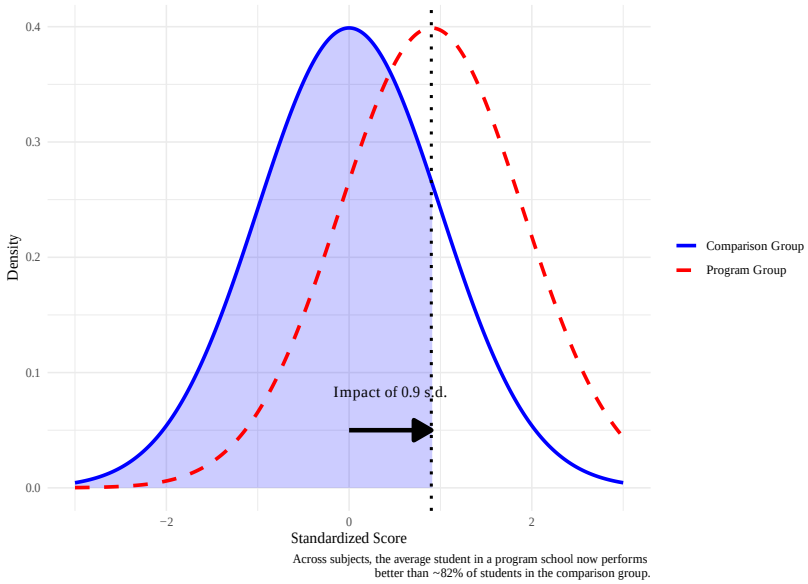
Notes: This figure confirms that the program and comparison schools exhibited similar levels and similar trends in exam scores before the program started. It presents, for the 138 program schools and 138 comparison schools, the yearly average end-of-primary exam scores over a period of seven years before the program launched. There is no data for the 2019-20 school year when the exam was canceled because of the Covid-19 pandemic (the figure connects the 2018-19 and 2020-21 values with a straight line). Shaded areas indicate 95 percent error bands.

Table E1: *The program led to sizable improvements in student learning*

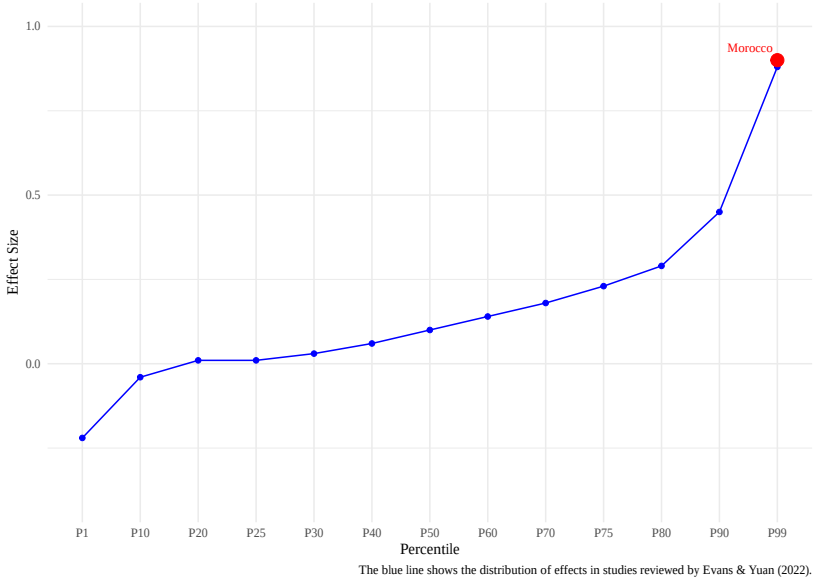
	Overall	By subject		
		Arabic	French	Math
	(1)	(2)	(3)	(4)
Panel A: Overall				
All students	0.90***	0.52***	1.30***	0.93***
	(0.03)	(0.05)	(0.05)	(0.04)
		[0.001]	[0.001]	[0.001]
Panel B: By gender				
Female	0.94***	0.51***	1.34***	0.94***
	(0.03)	(0.05)	(0.06)	(0.04)
	[0.001]	[0.001]	[0.001]	[0.001]
Male	0.87***	0.52***	1.23***	0.90***
	(0.04)	(0.06)	(0.07)	(0.05)
	[0.001]	[0.001]	[0.001]	[0.001]
Panel C: By baseline performance				
Bottom quartile	0.92***	0.63***	1.31***	0.83***
	(0.05)	(0.08)	(0.10)	(0.08)
	[0.001]	[0.001]	[0.001]	[0.001]
Top quartile	0.88***	0.41***	1.23***	0.97***
	(0.04)	(0.07)	(0.10)	(0.06)
	[0.001]	[0.001]	[0.001]	[0.001]

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 students' baseline learning level (as per the distribution within their enrolled grade level). ITT estimates follow equation (1) of the pre-analysis plan. Column (1) stacks the three subsamples and uses their respective test score as the outcome (Arabic, French, or mathematics). Standard errors are clustered at the matched-pair level and shown in parentheses. q-values are shown in brackets, using the Benjamini and Yekutieli (2001) correction for multiple hypothesis testing (MHT), and following Anderson (2008) to report the smallest level q at which the hypothesis of equal groups is rejected. * significant at 10%; ** significant at 5%; *** significant at 1%, before adjustment for MHT.

Figure E2: *The program's large impact can be interpreted more intuitively*



(a) *Interpretation with respect to the comparison group's test score distribution*



(b) *Comparison with program impacts observed in other countries*

Notes. These figures offer alternative interpretations of the program's overall impact of 0.9 standard deviations. The top figure illustrates how a positive shift of 0.9 standard deviations in the test score distribution means that, because of the program, the average student in a PSP school now performs better than 82 percent of their peers in the comparison schools that did not receive the program (assuming test scores are normally distributed). The bottom figure illustrates how, compared to the distribution of impacts found by Evans and Yuan (2022, Table 2) for 96 randomized controlled trials conducted in other low- and middle-income countries, the overall impact of 0.90 s.d. is larger than the 0.88 s.d. impact at the 99th percentile in Evans and Yuan that averages across reading and mathematics.

References

Anderson, M.L., 2008. Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. Journal of the American Statistical Association 103, 1481–1495. doi:10.1198/016214508000000841.

Benjamini, Y., Yekutieli, D., 2001. The Control of the False Discovery Rate in Multiple Testing under Dependency. The Annals of Statistics 29, 1165–1188.

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

GEEAP, 2023. Cost-Effective Approaches to Improve Global Learning: What Does Recent Evidence Tell Us Are Smart Buys for Improving Learning in Low- and Middle-Income Countries? Technical Report. The World Bank. Washington, D.C. URL: https://thedocs.worldbank.org/en/doc/231d98251cf326922518be0cbe306fdc-0200022023/related/GEEAP-Report-Smart-Buys-2023-final.pdf.

Contents

References **8**

1 Introduction **10**

2 Setting **12**

2.1 Public education in Morocco 12

2.2 The Pioneer School Program 13

3 Research methods **14**

3.1 Data 14

3.2 Sampling 15

3.3 Analytical strategy 16

3.4 Implementation quality 18

4 Results **18**

4.1 Main results 18

4.2 Secondary results 19

4.3 Exploring heterogeneity in effects 19

4.4 Effects on student dropout 20

5 Conclusion **21**

References **28**

Online Appendix **31**

A Additional tables **31**

B Sampling of schools **35**

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, 2019a). 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 the extent of learning that takes place in these schools has been disappointing in many, and perhaps most, of these countries. This raises the question of what can be done to increase student learning 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, 2023) identified "great buy" interventions, which are highly cost-effective and are supported by a strong body of evidence. However, consistently, such promising interventions have been found to be less effective (or even detrimental) if implemented without researcher oversight, once substantial external supports are removed, and when responsibilities are transferred from a non-governmental organization to the government (Vivalt, 2020; Allcott, 2015). This observation holds for the above-mentioned "great buy" recommendations, including programs involving scripted lesson plans (Kerwin and Thornton, 2021; Piper and Dubeck, 2024) and programs promoting that teachers target their instruction to a child's learning level (Banerjee et al., 2017; Duflo et al., 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 "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). It complements these program components by assigning some primary school teachers to specialize in teaching Arabic, French, and mathematics and by recognizing schools that successfully participate in the program (with a system of quality certification). The program was launched in 626 of Morocco's public primary schools in September 2023, and this study evaluates the program's impact on student learning over the program's first year. In secondary analyses, we also report on the program's impact on 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 seven years before the program was implemented support this "parallel trends" assumption.

The estimated intent-to-treat (ITT) impacts of the PSP program indicate extremely large impacts. 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 impact is 0.52 s.d. in Arabic, 1.30 s.d. in French, and 0.93 s.d. in mathematics. The estimated impacts are very similar for female and male students, with slightly larger impacts for female students in French and mathematics and for the overall score. Disaggregating by baseline student performance (within each grade level), the impacts on Arabic are higher for students in the bottom 25 percent of students in a given grade's baseline distribution (0.63 s.d.) than for students in the top 25 percent of that distribution (0.41 s.d.), while the impacts on mathematics are lower (but still extremely high) for students in the bottom 25 percent (0.83 s.d.) than for students in the top 25 percent (0.97 s.d.). For French, the impacts are slightly higher for students in the bottom 25 percent of students in that distribution (1.31 s.d.) than for students in the top 25 percent (1.23 s.d.).

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). Compared to other programs in low- and middle-income countries, the overall impact of 0.90 s.d. is within the top 1 percent of impacts on student learning in mathematics and reading, based on a systematic review of program impacts conducted by Evans and Yuan (2022). In addition, the 0.52 s.d. impact for Arabic is slightly higher than the 0.50 s.d. impact at the 90th percentile of Evans and Yuan for reading, and the 1.30 s.d. impact for French is much higher than the 0.90 s.d. impact at the 99th percentile of Evans and Yuan for reading. Finally, the impact of 0.93 for

mathematics is higher than the 0.80 s.d. impact at the 99th percentile of Evans and Yuan for mathematics.

The rest of this report is organized as follows. The following section provides an overview of Morocco's system of public education and then describes the Pioneer School Program. Section 3 explains the research methods used, starting with a description of the data and the selection of the schools used in the study, then turning to the analytical strategy, and finishing with a discussion of the quality of the implementation of the program. Section 4 presents the main results for the impact of the program on student learning, checks for heterogeneity of these program impacts, and also examines the impact on student dropout. The final section summarizes the conclusions and describes our next steps.

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 challenges surrounding low persistence rates into secondary school and with low student performance. Nearly 2.3 million children dropped out of primary and lower-secondary schools between 2008 and 2017 (World Bank, 2019b), 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 "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 2023, and this report evaluates the program's impact over the program's first year.[1] The Ministry considers the program its flagship intervention in primary schools, and it intends to scale the program to another 2,000 schools in the 2024-25 school year, covering approximately 30 percent of the students in the country. The remaining schools are expected to be reached in 2025-26.

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.[2] 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 teachers deliver through slide decks they receive from the Ministry.

The program enables these two components by allowing schools to assign some of their primary school teachers to specialize in teaching Arabic, French, and mathematics and by recognizing schools that successfully participate in the program through a system of quality certification (the schools thus obtain a "Pioneer School" label). Upon their successful participation in the program, the teaching staff of program schools also receive an individual one-time bonus of MAD 10,000 (approximately USD 1,000). All of the program implementers are regular teachers and inspectors commonly assigned to the program schools, with no additional staff being assigned to schools or NGO involvement in the program.

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

[2] 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) are not responsible for the program's implementation.

3 Research methods

Our research design and methods follow a registered pre-analysis plan.[3] This plan prespecified all the analyses, tables, and figures that we present in the main text of this article.

3.1 Data

This study uses three sources of data. The first is administrative data for all of Morocco's public primary schools. The second is baseline and endline assessment data that measure students' performance in Arabic, French, and mathematics. The third data source is process monitoring data collected during school visits (the latter dataset has yet to be shared with the research team and is not used in this report).

Morocco's Ministry of National Education Preschools and Sports granted the authors access to retrospective administrative data on all 21,077 public primary schools in Morocco over a 7-year period, from 2015 to 2022. The administrative data included more than 248 variables measuring student performance and socioeconomic status for 24 million students, as well as many variables on the characteristics of approximately the 104,000 teachers in these schools, and of the 21,077 schools themselves.

This study's main outcome of interest is student performance on academic assessments in Arabic, French, and mathematics. Baseline assessments for each of these three subjects were administered in September of 2023, and endline assessments were administered in June and July of 2024. The assessments record students' responses to all test questions (or test "items"). For each subject, a two-parameter logistic (2PL) item response theory (IRT) model was estimated to aggregate these responses and generate continuous estimates of student ability, using overlapping items (or "anchors") to map test scores onto common scales (across grades as well as across the baseline and endline data collection rounds).[4] In doing so, differential item functioning was investigated, and only test questions with stable item parameters were retained. Each subject's continuous test score was standardized by subtracting the mean, and then dividing by the standard deviation, of the distribution of the test scores at endline of the students in the matched (non-PSP) schools.

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

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

3.2 Sampling

This study's sample of schools was constructed in four steps, using administrative data on the universe of government schools, which includes the 626 PSP schools in Morocco. First, using the "post-double-selection" (PDS) methodology (Belloni et al., 2012, 2011, 2014, 2016) 16 variables were identified that were predictive of either baseline test scores or participation in the program. Second, using these 16 selected variables, Mahalanobis nearest-neighbor matching (without replacement) was used to identify a matched (non-PSP) school for each of the 626 PSP schools. Third, stratifying by region, 150 of the PSP schools (along with their 150 non-PSP matches) were randomly selected from the 626 PSP schools (and their matched non-PSP schools). Fourth, of the 300 sampled schools (150 PSP and their matched 150 non-PSP schools), baseline data collection could not be conducted in 19 schools. In addition, five of the remaining 281 schools (three PSP schools and two non-PSP schools) lost their matched schools due to this reduction of the sample to 281 schools. Dropping these five schools without a matched school results in the study's effective sample of 276 schools (138 PSP and 138 matched non-PSP). Appendix B provides additional technical details on the sampling of schools.

We compare the study's sample of program schools with the remaining primary schools in the country, other program schools, and the set of matched comparison schools. Appendix Table A1 shows evidence of positive selection of schools into the program and program schools into the study, but there are no meaningful differences between the program schools sampled for the study and their matched comparison schools. Compared to other schools in the country, the sampled schools are larger, predominantly urban, and serve a population of students that is less poor as per students' eligibility for social transfer programs (Columns 1 to 3).[5] At the same time, their average primary school leaving exam scores are slightly lower. The sample of program schools is more representative of other program schools that were not sampled, but it is also more urban and serves students who are less poor (Columns 4 to 6). Yet, as can be expected from the study's matching procedure, there are no notable differences between the program schools included in the study and their matched comparison schools (Columns 7 and 8).

[5]Our school-level, administrative data includes the percentage of students eligible for social safety net programs. We include them as a proxy for poverty levels. We report on the share of students qualifying for a conditional cash transfer program called "*Tayssir*". We also include a broader measure of eligibility for social safety net programs by reporting on the share of students qualifying for *Tayssir*, housing subsidies, and a program that offers free school bags. We approximate the percentage by dividing a school's total number of eligible students by the total number of enrolled students and capping the percentage at 100 (students can be eligible for more than one program). Our administrative data is for the 2021-22 school year; *Tayssir* ended in December 2023.

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 difference in the attrition rate of students between the PSP and non-PSP schools.

Table 1 provides information on four student-level variables at baseline and attrition rates, separately for the Arabic, French, and mathematics assessments. It also presents differences in these variables between the PSP and non-PSP schools in column (4). Of the 15 differences calculated (5 variables for 3 different assessments), none is significant at the 5 percent level, and only two are significant at the 10 percent level, which is about what one would expect by random chance. Moreover, for each of the three assessments, a joint test of the significance of all five variables fails to reject the null hypothesis of no significant differences for all five variables (p-values range from 0.32 to 0.91). Appendix Table A2 provides analogous information and significance tests for the students who remained in the sample at baseline, and the findings are very similar. Overall, the students in the PSP and non-PSP schools are very similar in terms of their observable characteristics.

3.3 Analytical strategy

This report presents estimates of the intent-to-treat effect of the PSP intervention on the outcomes of interest using a strategy that combines difference-in-differences with matching methods, which compares changes over time in the outcome variables of interest for primary schools that were assigned to receive that intervention with the same changes of their matched comparison schools. For all outcomes, the following empirical 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 outcome Y for student i in grade g in school s in matched pair j in period t. $TREAT_{sj}$ is the assignment of school s to receive the education intervention, and $POST_t$ is a dummy variable equal to 1 if the outcome Y is measured at endline (i.e., after the

intervention was rolled out) and 0 if measured at baseline (i.e., before the intervention was rolled out).

The $\gamma_{i,g,s,j}$ term is a student fixed effect, which measures the impacts on Y_{igsjt} of all student characteristics, observed or unobserved, that do not change over time. However, it is possible that the impacts of some student characteristics change over time, so a vector of observed student characteristics, denoted by X_{igsj}, is added and multiplied by $POST_t$. The corresponding vector of coefficients for this interaction term, denoted by δ, measures the change in the impacts of these observed student characteristics over time. Put another way, δ measures the contributions of observed student characteristics to changes in the outcome variables over time. Since nearly all students stay in the same school between baseline and endline data collection, student fixed effects also include school fixed effects, and so the X_{igsj} variables can also include observed school variables. The student variables available in the dataset are students' gender and whether the student was eligible for a social safety net measure (measured at baseline). The school characteristics are school-level average test scores, region, average class size, and the student-to-teacher ratio (all measured at baseline). A LASSO is used to determine in a data-driven way which $X_{igsj} \times POST_t$ interactions to include in the specification.[6]

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. Intuitively, the inclusion of these fixed effects accounts for how much, on average, a treatment student's matched-pair comparison peer (who is enrolled in the same grade level) learned between baseline and endline.

The coefficient of interest is β, which captures the intent-to-treat (ITT) effect of assignment to the program. In estimating ITT effects, we allow for the (plausible) possibility that the program was not fully implemented in at least some of the 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, taking into account a prespecified hierarchy of research hypotheses. In additional, exploratory analyses of heterogeneous treatment effects, we use the Benjamini and Yekutieli (2001) correction, following Anderson (2008).

For β in Equation (1) to estimate the causal impact of the PSP program on the outcome(s) of interest, the identifying assumption implies that, conditional on the "control" variables in that equation, the average trend in the outcome(s) over time in the PSP schools would

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

have been the same as the average trend in the matched non-PSP schools if the PSP had not been implemented in the PSP schools. While this assumption cannot be directly tested, one piece of supportive evidence is shown in Figure 1. It shows that average scores on the end-of-primary exams in these two sets of schools are very similar in the seven school years before the PSP was implemented.

3.4 Implementation quality

As stated in the pre-analysis plan, accessing process monitoring data is required to complete the study's preregistered analyses and publish its final report. We are grateful for the Ministry's commitment to sharing this data with us.

4 Results

4.1 Main results

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 extremely large impacts. Averaging over all three subjects, the impact of the PSP is 0.90 standard deviations (s.d.), which to the authors' knowledge, is larger than any impact ever estimated for a program implemented by a government (as distinct from programs implemented by non-governmental organizations). The impacts are also very high for each of the three subjects, ranging from a very high impact of 0.52 s.d. for Arabic to an astoundingly large impact of 1.30 s.d. for French. Compared to the distribution of impacts found by Evans and Yuan (2022, Table 2) for 96 randomized evaluations conducted in low- and middle-income countries, the overall impact of 0.90 s.d. is larger than the 0.88 s.d. impact at the 99th percentile in Evans and Yuan that averages across reading and mathematics. The 0.52 s.d. impact for Arabic is slightly higher than the 0.50 s.d. impact at the 90th percentile of Evans and Yuan for reading, and the 1.30 s.d. impact for French is much higher than the 0.90 s.d. impact at the 99th percentile of Evans and Yuan for reading. Finally, the impact of 0.93 for mathematics is much higher than the 0.80 s.d. impact at the 99th percentile of Evans and Yuan for mathematics.

4.2 Secondary results

The pre-analysis plan for this study also included secondary analyses that separate estimates by gender and by students' (within-grade) baseline performance. The impacts disaggregated by gender in Panel B of Table 2 show very similar impacts for male and female students, with slightly larger impacts for female students in French and mathematics, and for the overall score. 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 for both male and female students and for students at the top and the bottom of the distribution of baseline student performance.

In additional pre-specified analyses in the Appendix, we confirm the estimated effects are robust to using alternative outcome measures. Panels A and B of Appendix Table A3, show the program led to substantial improvements in reading fluency (an increase of 7.2 and 8.4 words read correctly per minute in Arabic and French, respectively). Panel C of Table A3 shows the effects in math are not concentrated only on the "calculation and numeracy" content domains many remedial programs focus on more strongly.[7] If anything, the program's effects on the "geometry, measures, and data" content domain are slightly larger (0.97 vs. 0.87 s.d.). Moreover, the impacts are similar for questions that require application and reasoning vs. questions that rely more heavily on rote learning (or "knowledge of procedures"). In Arabic and mathematics, the positive effects appear to be slightly larger for at-grade material; but both at-grade and below-grade content are affected positively, and the difference in effects is small.[8] Lastly, the program impacts in math are robust to giving equal weight to the content subdomains measured by the math assessment.

4.3 Exploring heterogeneity in effects

In addition to the above investigation of program impacts among girls and low-performing students, we also explored additional heterogeneity in treatment effects along a broader, pre-specified set of student and school characteristics. More specifically, to better

[7] For example, the "Teaching at the Right Level" program focuses more strongly on number recognition and basic arithmetic and de-emphasizes spatial skills. Relatedly, the "ASER" assessments do not measure spatial skills.

[8] We also planned to conduct a similar analysis for French. However, for students enrolled in grades 3, 5, and 6, all of the test questions capture material from below their grade level, and therefore, we did not perform this analysis by curricular grade level in French.

understand who benefited the most from the program, we used the causal forest approach developed by Athey and Imbens (2016) and Wager and Athey (2018).

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

Following Carlana et al. (2022) and Dinarte et al. (2024), we use the out-of-bag predictions to categorize students into the top and bottom half of predicted treatment effects (designated as "Weak" and "Strong" groups). 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 3 presents the results.

As shown in the top row, there is little heterogeneity in impacts, with the bottom and top half of (predicted) impacts differing by just 0.13 s.d. Panel A shows that female students, students in the top quartile at baseline, students in grades 5 and 6, and students with higher baseline scores are more likely to be among the 50 percent of students who experienced greater program impacts (the "Strong" group). Also, Panel B shows that schools with higher primary school exam scores experienced slightly larger impacts. Yet again, the difference in effects among the "Weak" and "Strong" groups is small, so being more prone to belonging to either one of these two groups does not translate into meaningfully large differences in program effects.[9]

4.4 Effects on student dropout

In Morocco, student dropout from public primary schools is concentrated in grade 6, at the end of the school year, and when students (fail to) transition from primary school to secondary school (Gazeaud and Ricard, 2024). So far, we only have data on whether the study's sample of students persisted in school throughout the same school year. In Table 4, we present the results of regressing a student's enrollment status at the end of the school year on a program-school indicator and matched pair-by-grade fixed effects. At the bottom

[9]Relatedly, recall that there is no notable difference in the program effects by students' baseline performance (Table 2, Panel C).

of the table, we also present the average percentage of students dropping out throughout the school year.

Table 4 shows that virtually all students remained enrolled, with less than 1 percent of students dropping out throughout the school year. In additional analyses (not shown in the table), we find the percentage of students dropping out throughout the school year is not associated with students' grade level (including whether students are in grade 6) or any other observable student or school characteristic, including student test scores, or whether the school location is rural. Perhaps unsurprisingly, we also find that whether students remained enrolled throughout the school year was not affected by the program, both overall (Panel A) and if we split the sample by gender (Panel B).[10]

We caution that these preliminary results may be better interpreted as "continued enrollment throughout the school year" (instead of dropout). We are more interested in whether students continued to be enrolled in school in the *following* school year (2024-25) and especially whether grade-6 students continued their schooling in secondary school. We are currently in the process of obtaining the necessary data to answer this question; we thank the Ministry for its support and look forward to updating our results accordingly. Until then, we strongly advise against interpreting the above findings as evidence that the program has no impact on dropout rates, nor should the findings be interpreted as indicating that there is virtually no student dropout among Morocco's primary school students.

5 Conclusion

To address the problem of very low learning levels in public primary schools and reduce student dropout, the Moroccan Ministry of National Education Preschool and Sports turned to global evidence of best-practice interventions that successfully address the global learning crisis that plagues many low- and middle-income countries. Its "Pioneer School Program" is inspired by the Global Education Evidence Advisory Panel (GEEAP, 2023) and combines two of its "great buy" recommendations on how to address the learning crisis. The program's main components consist of structured pedagogy (including scripted lessons) and targeted remediation (by student learning level, not grade). It complements these program components by assigning some primary school teachers to specialize in teaching Arabic, French, and mathematics and by recognizing schools that successfully participate in the program (with a system of quality certification). While promising, similar

[10]Our registered pre-analysis plan also included an exploration of heterogeneous treatment effects among students of high (vs. low) risk of dropping out. Since we are unable to predict which students are more (or less) likely to drop out, we removed this analysis from Table 4.

education interventions are difficult to scale, and impacts in government schools are often muted. Thus, whether the program would indeed lead to improved learning levels was an open question.

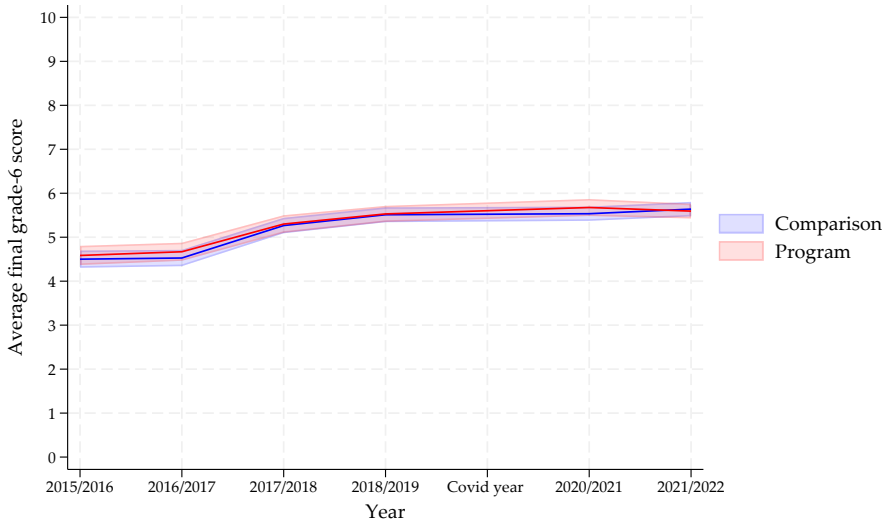
Our study finds that, after one year, the program led to very large improvements in student learning. Averaging over all three subjects, the intent-to-treat (ITT) impact of the PSP is 0.90 standard deviations (s.d.) of the distribution of the scores of non-PSP schools at endline. By subject, the program's impact is 0.52 s.d. in Arabic, 1.30 s.d. in French, and 0.93 s.d. in mathematics. We find that similarly large, positive impacts also hold for female students and for students whose learning levels were farther behind when the program launched. These results are robust to alternative measures of student learning, including for content and cognitive subdomains not explicitly targeted by the program.

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). The program's average effect exceeds the 99th percentile of treatment effects of other education interventions in low- and middle-income countries. These findings suggest that an integrated intervention can achieve transformative impacts on student learning, including when a program is implemented at scale by a government without support from external organizations.

Our next steps consist of adding the data on implementation fidelity and program take-up to our analysis and updating our analysis of effects on student dropout rates with data that follows the study's sample of students into the next school year (including for grade-6 students, whether they enrolled in secondary school). We will also start to disseminate the study's results, including at the upcoming *Forum National de l'Enseignant* in Rabat, and prepare a submission of our findings to an academic journal.

We are eager to build on this successful collaboration with the Ministry and to engage in similar research studies (with the J-PAL-affiliated Morocco Innovation and Evaluation Lab). If there is interest from the Ministry, this may also include randomized trials to investigate which of the program components drive the success of the PSP program.

Figure 1: *Parallel trends*



Notes: This figure confirms that the program and comparison schools exhibited similar levels and similar trends in exam scores before the program started. It presents, for the 138 program schools and 138 comparison schools, the yearly average end-of-primary exam scores over a period of seven years before the program launched. There is no data for the 2019-20 school year when the exam was canceled because of the Covid-19 pandemic (the figure connects the 2018-19 and 2020-21 values with a straight line). Shaded areas indicate 95 percent error bands.

Table 1: *Sample of students and balance tests*

	Sample size		Balance	
	Comparison	Program	Comparison	Difference
	(1)	(2)	(3)	(4)
Panel A: Arabic				
% Female	3816	3890	49.79	-1.08
			[50.01]	(1.36)
% Ever repeated	3458	3517	12.98	-0.14
			[33.62]	(0.88)
% Ever qualified for Tayssir	3816	3890	22.38	0.15
			[41.68]	(1.06)
Baseline test score	3816	3890	-0.37	-0.02
			[0.95]	(0.05)
% Attrited	3816	3890	12.40	-2.77
			[32.96]	(2.25)
Joint F-test (p-value)				0.85
Panel B: French				
% Female	3719	3848	50.04	0.03
			[50.01]	(1.32)
% Ever repeated	3361	3483	12.88	0.72
			[33.51]	(0.93)
% Ever qualified for Tayssir	3719	3848	23.34	-0.93
			[42.30]	(1.02)
Baseline test score	3657	3800	0.19	-0.11*
			[1.02]	(0.06)
% Attrited	3719	3848	10.49	-2.08
			[30.64]	(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)
% Ever qualified for Tayssir	3729	3844	21.99	1.31
			[41.42]	(0.96)
Baseline test score	3729	3844	-0.60	-0.06*
			[0.82]	(0.03)
% Attrited	3729	3844	7.51	-1.38
			[26.36]	(1.21)
Joint F-test (p-value)				0.23

Notes. Sample and unit of observation: 20,784 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. "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)
		[0.001]	[0.001]	[0.001]
Panel B: By gender				
Female	0.94***	0.51***	1.34***	0.94***
	(0.03)	(0.05)	(0.06)	(0.04)
	[0.001]	[0.001]	[0.001]	[0.001]
Male	0.87***	0.52***	1.23***	0.90***
	(0.04)	(0.06)	(0.07)	(0.05)
	[0.001]	[0.001]	[0.001]	[0.001]
Panel C: By baseline performance				
Bottom quartile	0.92***	0.63***	1.31***	0.83***
	(0.05)	(0.08)	(0.10)	(0.08)
	[0.001]	[0.001]	[0.001]	[0.001]
Top quartile	0.88***	0.41***	1.23***	0.97***
	(0.04)	(0.07)	(0.10)	(0.06)
	[0.001]	[0.001]	[0.001]	[0.001]

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 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. Column (1) stacks the three subsamples and uses their respective test score as the outcome (Arabic, French, or mathematics). Standard errors are clustered at the matched-pair level and shown in parentheses. q-values are shown in brackets, using the Benjamini and Yekutieli (2001) correction for multiple hypothesis testing (MHT), and following Anderson (2008) to report the smallest level q at which the hypothesis of equal groups is rejected. * significant at 10%; ** significant at 5%; *** significant at 1%, before adjustment for MHT.

Table 3: *Exploring conditional average treatment effects on student learning*

	Weak Group	Strong Group	Diff.	MHT q-value
	(1)	(2)	(3)	(4)
ITT effect	0.857	0.991	0.134	
Panel A: Student characteristics				
Female	44.103	54.561	10.458	0.001
Top quartile	17.426	32.560	15.134	0.001
Middle two quartiles	48.563	50.729	2.166	0.272
Bottom quartile	34.011	16.711	-17.300	0.001
Grade 1	12.859	15.349	2.489	0.010
Grade 2	21.716	10.601	-11.115	0.001
Grade 3	22.695	11.655	-11.040	0.001
Grade 4	12.976	21.873	8.897	0.001
Grade 5	15.244	19.447	4.203	0.001
Grade 6	14.509	21.075	6.566	0.001
Baseline score	-0.595	0.047	0.642	0.001
Ever repeated	13.392	12.400	-0.991	0.222
Ever qualified for Tayssir	25.271	24.407	-0.865	0.986
Panel B: School characteristics				
Number of teachers	21.507	17.729	-3.777	0.918
Urban (vs rural)	73.153	67.344	-5.809	1.000
Regional development	58.186	64.055	5.869	1.000
Total enrollment (2021/2022)	598.802	475.770	-123.032	0.712
Female students (percentage, 2021/2022)	48.116	48.257	0.142	1.000
Tayssir beneficiary (percentage, 2021/2022)	15.283	20.400	5.116	1.000
Safety net beneficiary (percentage, 2021/2022)	15.335	20.560	5.224	1.000
Average grade-6 score (2021/2022)	5.312	5.891	0.580	0.001

Notes. This table explores heterogeneity in treatment effects. We stack the three subsamples and use their respective test score as the outcome (Arabic, French, or mathematics). The sample consists of 18,789 students with non-missing background information. "Strong group" refers to subgroups whose conditional average treatment effect (CATE) is above the median of all CATEs when switching to the treatment (and below the median for the "Weak group"). A positive number in the Difference column indicates that the average covariate value for the "Strong group" is higher. q-values for the difference between groups are shown in the fourth column, clustering standard errors for student characteristics at the matched-pair level, using the Benjamini and Yekutieli (2001) correction for multiple hypothesis testing (MHT), and following Anderson (2008) to report the smallest level q at which the hypothesis of equal groups is rejected. 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).

Table 4: *Intent-to-treat effects on students' enrollment status at the end of the 2023-24 school year*

	Student dropout (%)
Panel A: Overall	
All students	0.0600
	(0.16)
Panel B: By gender	
Female	0.0780
	(0.18)
	[1.000]
Male	0.0412
	(0.22)
	[1.000]
Comparison group mean	0.68

Notes. This table reports on the program's intent-to-treat (ITT) effects on students' enrollment status at the end of the 2023-24 school year, among all students (Panel A) and by gender (Panel B). Data on whether students continued their schooling in the 2024-25 school year are not yet available (results are forthcoming). Estimates are yielded by regressing an indicator of whether a student had dropped out by the end of the 2023-24 school year on treatment status and grade-by-matched pair fixed effects. Unlike what was pre-specified, we do not show a separate panel of impacts among students with a high (vs. low) predicted dropout risk, as the available data is unable to predict a student's enrollment status at the end of the 2023-24 school year. Standard errors are clustered at the matched-pair level and shown in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%. Westfall-Young stepdown adjusted p-values are shown in brackets.

References

Allcott, H., 2015. Site Selection Bias in Program Evaluation. The Quarterly Journal of Economics 130, 1117–1165. doi:10.1093/qje/qjv015.

Anderson, M.L., 2008. Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. Journal of the American Statistical Association 103, 1481–1495. doi:10.1198/016214508000000841.

Angrist, N., Djankov, S., Goldberg, P.K., Patrinos, H.A., 2021. Measuring human capital using global learning data. Nature 592, 403–408. doi:10.1038/s41586-021-03323-7. publisher: Nature Publishing Group.

Athey, S., Imbens, G., 2016. Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences 113, 7353–7360. doi:10.1073/pnas.1510489113.

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

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

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

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

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

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

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

Benjamini, Y., Yekutieli, D., 2001. The Control of the False Discovery Rate in Multiple Testing under Dependency. The Annals of Statistics 29, 1165–1188.

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

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

Dinarte, L., Egana-delSol, P., Martinez A., C., Rojas A., C., 2024. When Emotion Regulation Matters: The Efficacy of Socio-Emotional Learning to Address School-Based Violence in Central America. doi:10.2139/ssrn.4741457.

Duflo, A., Kiessel, J., Lucas, A.M., 2024. Experimental Evidence on Four Policies to Increase Learning at Scale. The Economic Journal , ueae003doi:10.1093/ej/ueae003.

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

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

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

GEEAP, 2023. Cost-Effective Approaches to Improve Global Learning: What Does Recent Evidence Tell Us Are Smart Buys for Improving Learning in Low- and Middle-Income Countries? Technical Report. The World Bank. Washington, D.C. URL: https://thedocs.worldbank.org/en/doc/231d98251cf326922518be0cbe306fdc-0200022023/related/GEEAP-Report-Smart-Buys-2023-final.pdf.

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

Kerwin, J.T., Thornton, R.L., 2021. Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures. The Review of Economics and Statistics 103, 251–264. URL: https://doi.org/10.1162/rest_a_00911, doi:10.1162/rest_a_00911.

Mansouri, Z., Moumine, M.E.A., 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, 9. URL: http://ijere.iaescore.com/index.php/IJERE/article/view/6341, doi:10.11591/ijere.v6i1.6341.

Piper, B., Dubeck, M.M., 2024. Responding to the learning crisis: Structured pedagogy in sub-Saharan Africa. International Journal of Educational Development 109, 103095. doi:10.1016/j.ijedudev.2024.103095.

UIS-UNESCO, 2024. UIS statistics. URL: http://data.uis.unesco.org/.

Vivalt, E., 2020. How Much Can We Generalize From Impact Evaluations? Journal of the European Economic Association , 1–45URL: https://academic.oup.com/jeea/advance-article/doi/10.1093/jeea/jvaa019/5908781, doi:10.1093/jeea/jvaa019.

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

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

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

World Bank, 2019a. Ending Learning Poverty: What Will It Take? Technical Report. The World Bank. Washington, D.C. URL: https://hdl.handle.net/10986/32553. publisher: World Bank, Washington, DC.

World Bank, 2019b. Morocco - Education Support Program Project. Project Appraisal Document PAD3157. The World Bank. Washington, D.C. URL: https://documents.worldbank.org/en/publication/documents-reports/documentdetail/908441561140203130/Morocco-Education-Support-Program-Project.

Online Appendix

A Additional tables

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]	(1.76)
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]	(75.26)
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]	(1.13)
Tayssir beneficiary (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]	(7.09)
Safety net beneficiary (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]	(7.09)
Average final grade-6 score (2021/2022)	5.62	5.74	-0.13*	5.59	5.67	-0.07	5.65	0.01
	[1.00]	[1.14]	(0.07)	[1.01]	[1.05]	(0.10)	[0.99]	(0.21)

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). 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)
% Ever 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)
% Ever 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)
% Ever qualified for Tayssir	3449	3617	22.09	1.01
			[41.49]	(1.02)
Baseline test score	3449	3617	-0.59	-0.06*
			[0.82]	(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%.

Table A3: *Effects on subdomains of student learning*

	ITT Effect
Panel A: Arabic	
Words read correctly / minute	7.24***
	(0.71)
At grade	0.54***
	(0.05)
Below grade	0.39***
	(0.05)
Panel B: French	
Words read correctly / minute	8.43***
	(2.48)
Panel C: Mathematics	
At grade	0.97***
	(0.04)
Below grade	0.87***
	(0.04)
Calculation and numeracy	0.87***
	(0.04)
Geometry, measures, and data	0.97***
	(0.04)
Knowing	0.87***
	(0.04)
Applying and reasoning	0.83***
	(0.04)
Equal weights to content domains	0.96***
	(0.04)

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). Standard errors are clustered at the matched-pair level and shown in parentheses. * significant at 10%; ** significant at 5%; *** significant at 1%.

B 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 were able to implement 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 used is post-double selection LASSO (PDSLASSO) used for estimating structural parameters in linear models with many controls (Belloni et al., 2012, 2011, 2014, 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 (all 248 constructed variables in the database). 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, thereby enhancing the robustness of causal inference (Ahrens et al., 2019).

Using a school-level database with 248 variables, consisting of school-level characteristics and aggregated student and teacher characteristics in 8,034 primary schools (including the pre-selected program schools), the PDSLASSO implemented identified 16 variables to be used for matching.

Next, we implemented an iterative process of a calibrated Mahalanobis distance matching algorithm using the variables selected by the PDSLASSO, to match 626 untreated schools (from the universe of 7,408 untreated schools) to the 626 selected schools for treatment. 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 6 strata belonging to the 3-baseline outcome terciles in 2 regions with very few schools. As a result, we only stratified the school sampling by region for those 2 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, A., Hansen, C.B., Schaffer, M.E., 2019. PDSLASSO: Stata module for post-selection and post-regularization ols or iv estimation and inference. URL: https://ideas.repec.org//c/boc/bocode/s458459.html. Statistical Software Components, January.

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

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

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

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

Ham, D.W., Miratrix, L., 2024. Benefits and costs of matching prior to a difference in differences analysis when parallel trends does not hold. URL: http://arxiv.org/abs/2205.08644. arXiv.