Essays on the Economics of Education

Astrid Camille Pineda Taveras

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2025

Essays on the Economics of Education

Abstract

This dissertation is composed of three chapters. Each chapter is a paper in the field of economics of education. Together, they examine the impact of three educational interventions aimed at improving student outcomes, using experimental methods across diverse settings.

The first chapter estimates the impact of partially substituting instructional time for the use of computer-adaptive learning technology. To examine this, a field experiment was conducted randomizing the use of the software for two weekly hours among 38 9th grade classrooms. In addition to this, a tutoring component targeted to 20 percent of students was randomized among treated classrooms. Results show that partially substituting instructional time for the use of the adaptive platform led to large improvements in students' test scores. However, classrooms that received the tutoring component in addition to the adaptive learning software, performed worse than those that only received the latter. These differences seem partially explained by implementation challenges and lower use of the platform in tutored classrooms.

The second chapter examines short- and long-term impacts of attending urban and nonurban charter schools in Massachusetts. Using randomized admission lotteries, we find that urban charter schools increase test scores and college attendance and completion; whereas, nonurban charter schools decrease test scores but increase college attendance and graduation, even more so than urban charters. Results suggest that both urban and nonurban charter schools increase college preparation but via different paths. Urban charters increase AP and SAT outcomes, as well as the completion of MassCore, a state-recommended college-prep curriculum. Nonurban charters decrease AP-taking but have a large positive impact on MassCore completion. These findings

suggest that there is more than one path to college success, and test score impacts do not always predict longer-term gains.

Lastly, the third chapter examines the impact of a comprehensive reading intervention using a randomized field experiment. The intervention, implemented over two years in a sample of Haitian public schools, provided a combination of teacher training, coaching, supervision, and school supplies. By endline, 18 months after the start of the intervention, we find large positive effects on students' reading outcomes. Notably, our results suggest that a model relying on school principals for teacher coaching can be as effective as a costlier and more logistically challenging version involving external coaches. These findings highlight the potential for using existing school leadership to deliver scalable and sustainable pedagogical support in low-capacity settings.

Table of Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I thank my family. My mother has always been my biggest fan—eager to hear about every project I am working on and sharing them proudly with anyone willing to listen. Her unwavering support, unconditional love, and constant encouragement have carried me through every stage of this journey. I am also deeply grateful for the memory of my grandmother, Naná, who passed away last year. Her legacy of love, hard work, kindness, and integrity continues to inspire me to strive to become a better person every day.

I am extremely thankful for my wonderful advisors, Sarah Cohodes and Alex Eble. I feel so fortunate to have had kind, patient mentors who took the time to teach me how to be a better researcher and colleague through their guidance and example. I am also grateful to my committee members—Chris Neilson and Peter Bergman—for their insightful feedback and support throughout this process. I thank the colleagues who enriched our work together with both insight and enthusiasm: Jostin Kitmang, Carolina Lopez, Michela Carlana, Eliana La Ferrara, Juan Barón, and José Mola.

I am grateful to the professors and mentors from whom I have learned—and continue to learn—so much, and who offered valuable feedback throughout this process, especially Donald Green and Alejandro Ganimian.

I would also like to thank Ancell Shecker at the Ministry of Education of the Dominican Republic for her early and continued support. Her commitment to education and evidence-based policy made this work possible.

To my husband, Michael Alonso, whose love, patience, and grounding presence carried me through the hardest moments and made this journey possible.

I also thank the PEAR program for its financial support, and program coordinators Corinne Herlihy and Thomas Brock for their encouragement. Funding for the first chapter was provided by the Tinker Foundation, and implementation support was made possible through J-PAL.

The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B200017 to Teachers College, Columbia University. The opinions expressed are those of the author and do not represent the views of the Institute or the U.S. Department of Education.

Dedication

I dedicate this dissertation to my mother, Ana Cecilia Taveras, and to the memory of my grandmother, Ana Gloria Almánzar (Naná).

Introduction

A unifying theme across the three chapters of this dissertation is the evaluation of educational interventions aimed at improving student outcomes. While each chapter focuses on a different context—ranging from fragile state settings to middle-income and developed economies—they collectively provide insights into the types of approaches that can promote student success across varied educational systems.

All three chapters use rigorous empirical methods, particularly experimental designs, to assess the effectiveness of different educational strategies. The first chapter focuses on the Dominican Republic and evaluates the impact of integrating computer-adaptive learning software into classroom instruction. This chapter shows that, even when substituting for instructional time, the use of adaptive technology can lead to significant gains in student math performance. However, layering on a tutoring component led to smaller effects, underscoring the importance of implementation quality and the potential for unintended interactions between overlapping interventions.

The second chapter, set in Massachusetts, examines how different types of charter schools—urban and nonurban—affect students' academic outcomes. The findings reveal that urban charter schools significantly improve test scores and college outcomes, while nonurban charter schools, though not improving test scores, have strong positive effects on college attendance and completion, suggesting that standardized test scores may be a limited measure of students' long-term educational success.

The third chapter evaluates the effects of a structured reading intervention in Haitian public schools, where capacity constraints are substantial. The intervention combined teacher training, coaching, supervision, and provision of learning materials. The results demonstrate that both an intensive version involving external coaches and a leaner model relying on school principals led to meaningful gains in reading skills. These findings suggest that even in fragile settings, cost-effective and scalable models can be successful when they leverage existing school leadership and provide structured pedagogical support.

Taken together, the chapters show that well-designed educational interventions—whether tech-based, organizational, or pedagogical—can yield significant benefits for student learning and longer-term educational trajectories. They also emphasize the importance of adapting strategies to context and the potential for diverse pathways to educational success.

Chapter 1. Improving Student Learning through Adaptive Learning Platforms: Experimental Evidence from the Dominican Republic

1.1 Introduction

In 2018, less than 20 percent of students in low and middle-income countries (LMIC) were able to meet minimum proficiency levels in mathematics (World Bank, 2017a).[2] More recently, the learning crisis, already a pressing concern, has deepened as a result of the COVID-19 pandemic. Over the course of two years, from February 2020 to February 2022, schools around the world were closed for an average of seven months (UNESCO, 2021). To counter these negative impacts, various strategies have been explored, including in-person and online tutoring (Carlana and La Ferrara, 2021; Nickow et al., 2020), lesson delivery via phone calls (Angrist et al., 2020), and the use of Artificial Intelligence (AI) tools (Muralidharan et al., 2019), all of which have shown promising results in improving student outcomes. However, a pivotal question remains: how can school systems build more resilient methods of delivering instruction?

One promising approach lies in leveraging technology, particularly AI, to address the growing challenges associated with providing quality education at-scale. AI-powered tools offer the potential to accelerate learning recovery and transform education systems worldwide. Among these, computer-adaptive learning (CAL) platforms stand out as highly effective (Antoninis et al., 2023).

[1] Co-authored with Carolina Lopez.

[2] This knowledge threshold indicates that students below this level have not yet acquired basic math skills, including simple calculations with whole numbers, or interpreting bar graphs.

In classrooms with diverse learning levels, where teachers may struggle to address individual needs, CAL tools adapt lessons based on each student's responses, tailoring activities to match their abilities—all at a fraction of the cost of one-on-one tutoring. Additionally, they provide students with instant feedback and equip teachers with valuable data to track progress and performance.

In this study, we evaluate through a multi-stage randomized controlled trial (RCT) the impact of an innovative intervention combining the use of a CAL platform with group tutoring, with the goal of accelerating learning recovery among high school students in the Dominican Republic (DR). While the introduction of these new technologies may accelerate learning for all students within a classroom, it may not suffice to significantly narrow the substantial learning disparities that often exist among students. With this goal in mind, we combine the use of CAL with targeted group tutoring for students most in need. The study design and survey instruments allow us to generate evidence on whether the use of CAL, by itself, or when combined with targeted tutoring, can yield improvements in student outcomes.

In 2023, our intervention introduced the use of CAL as part of regular instruction for 9th graders, replacing two out of the seven weekly hours allocated for regular math instruction. Randomization was conducted at the classroom level (38 classrooms in total). The software chosen, ALEKS, developed by McGraw Hill, has been adapted to the DR math curriculum and was used in approximately 60 schools across the country, yielding satisfactory qualitative results.[3] All students in treated classrooms received free licenses to access this software. Additionally, within the selected classrooms, we randomized access to in-person tutoring by providing up to three weekly hours of tutoring to high-need students. This component aimed to explore complementarities between CAL and more traditional forms of instruction. Students participating in tutoring sessions had access to the platform, and tutors were instructed to use the platform data to monitor student progress, identify learning gaps, and plan their lessons. These additional lessons were conducted during

[3]The intervention was funded by the World Bank and implemented by the Ministry of Education during the years 2018-2023.

school hours otherwise dedicated to extracurricular activities.

We find that the use of the adaptive software during math classes improved student outcomes by 0.29 to 0.31 standard deviations by endline (roughly six months after the start of the intervention). These positive effects are robust across various model specifications and consistent across multiple knowledge domains, including algebra and geometry, as well as different difficulty levels, such as content below and at grade level. However, we do not find statistically significant differences between classes that received the tutoring component and the control group, suggesting that adding tutoring to the CAL component may have hindered the effectiveness of the latter.

Why did the two treatment arms have different impacts? While we do not find conclusive evidence explaining these negative differences, we observe that students selected for tutoring were less likely to use the adaptive platform (as were their classmates), exhibited lower levels of locus of control (i.e., belief in their ability to influence their own outcomes), and experienced lower levels of math anxiety compared to the CAL-only group. Additionally, we encountered significant implementation challenges that may have affected the quality and consistency of the tutoring sessions, including high student absenteeism. We also find evidence indicating that the tutoring may not have been well-targeted, as students selected by teachers to receive tutoring were not among the lowest performers at baseline, as was intended by the program.

Our paper contributes to the growing literature on remedial education, particularly tutoring. Tutoring has been shown to be one of the most effective interventions for improving learning outcomes. A recent meta-analysis by Nickow et al. (2020) reported an average effect size of 0.37 standard deviations for tutoring interventions, with one-to-one tutoring showing the largest effects. However, the scalability of this approach is hindered by high costs and the challenge of recruiting qualified tutors (Kraft and Falken, 2020; Nickow et al., 2020; Kraft et al., 2024), and the longstanding issue of finding scalable alternatives that are as effective, known as the "2-sigma problem" (Bloom, 1984), remains a significant concern. Recent evidence, however, suggests that small-group tutoring (up to five students per tutor) can achieve similar results at lower costs (Augustine et al., 2016; Guryan et al., 2023; Kraft and Falken, 2020; Nickow et al., 2020; Kraft et al., 2024). Our study contributes to this literature by examining the combination of small-group tutoring with the use of adaptive learning technologies in school, which could enhance the effectiveness of traditional group tutoring by leveraging CAL data to better address specific learning deficiencies.

Despite limited research on AI-driven education interventions (Miao et al., 2021), increasing access to computer-adaptive technologies has spurred a growing body of research (Angel-Urdinola et al., 2023; Guryan et al., 2023; Muralidharan et al., 2019).[4] Most existing studies have evaluated CAL in after-school settings or combined with tutoring (Muralidharan et al., 2019; Guryan et al., 2023), leaving open questions about its effectiveness when integrated into regular school schedules, or implemented alone (i.e. without a tutoring component) (Muralidharan et al., 2019). Additionally, because of the setting in which these studies have taken place, the impact of CAL alone versus additional instructional time remains difficult to disentangle (Glewwe and Muralidharan, 2016; Muralidharan et al., 2019). To our knowledge, this study is among the first to test the substitution of regular math instruction with the use of an adaptive learning platform (Muralidharan and Singh, 2023). We replace two out of seven weekly math instruction hours with CAL sessions, offering insight into its efficacy as part of the standard school curriculum.

While most of the literature on teacher effectiveness has focused on measuring teacher value-added (Chetty et al., 2014a; Rockoff, 2004), identifying causal links between teacher effectiveness and student outcomes (Chetty et al., 2014b; Gershenson, 2016; Rivkin et al., 2005; Rockoff, 2004), or examining teacher characteristics (i.e. years of education or experience) associated with high value-added (Chingos and Peterson, 2011; Nye et al., 2004). It is less understood *how* to improve teacher effectiveness, and particularly, whether the introduction of new technology in the classroom can

[4] We focus on computer-adaptive learning, which personalizes content to students' actual learning levels, as opposed to traditional computer-assisted instruction, which reviews only grade-appropriate content. Studies of the latter include Angrist and Lavy (2002), Barrow et al. (2009), and Linden et al. (2003).

improve teacher effectiveness. Taylor (2015) evaluates the impact of computer-aided instruction on teacher effectiveness and finds that its introduction led to changes in how teachers' allocated class time. Using student-reported data, we explore whether the integration of an adaptive platform in class leads to changes in teachers' time use and practices, as this is one potential mechanism by which the intervention may have an impact on teacher effectiveness and student performance.[5] We do not observe changes in teachers' practices or allocation of time, as reported by students, suggesting that the effects observed may be attributed to the direct substitution of teachers' instruction for the use of the learning platform, rather than changes to how teachers allocated time and effort across tasks outside this period.

The deepening of the learning crisis due to COVID-19 highlights the need for more resilient education systems that can better withstand disruptions. Policymakers could invest in building infrastructure and capacity for blended learning models that integrate digital tools like CAL into teachers' work, ensuring continuity of learning in times of crisis. Given the high costs and logistical challenges associated with traditional tutoring, this study suggests that AI-driven adaptive learning platforms could serve as a scalable alternative for improving student outcomes. Expanding access to such technologies can provide affordable and effective learning support at scale, particularly in resource-constrained settings.

This paper proceeds as follows. In Section 2, we describe in detail the context, the intervention and its components. In Sections 3 and 4, we discuss the study design and describe our data. In Section 5, we present the empirical framework and the results of our analysis. Lastly, in Section 6, we conclude.

[5] It is important to note that the context of this study is different from that in Taylor (2015). While Taylor (2015) examines experiments that took place in the US, this work takes place in the Dominican Republic, one of the lowest performing countries in international assessments. For instance, when looking at the 2018 PISA Math results, the US ranked in the 9th place while the DR ranked 78th (out of 79 countries).

1.2 Intervention and Study Design

1.2.1 Background

The study takes place in the Dominican Republic (DR), a country facing substantial gaps in key knowledge areas such as mathematics and reading comprehension. In 2022, the DR ranked third to last in math on the OECD's Programme for International Student Assessment (PISA), with 92 percent of students classifying as low-performing (OECD, 2023a,b). The Dominican education system is divided into primary (grades 1-6), lower secondary (grades 7-9) and upper secondary (grades 10-12) levels, with students typically entering lower secondary around age 12. The school year in the DR is organized over two academic semesters, and usually begins in late August/early September and ends in late June.[6]

The transition to secondary schooling represents a critical juncture in the academic trajectory of Dominican students. Previous research indicates that up to 40 percent of students drop out between 8th and 12th grade (IDEICE, 2017). These high dropout rates, coupled with notably low performance levels, underscore the urgency of implementing remedial interventions to improve student learning outcomes and retention. This challenge is not unique to the DR. Across Latin America and the Caribbean (LAC), countries face similar educational obstacles. The LAC region ranks in the bottom half of the global scale for educational quality in mathematics (IDB, 2023): the difference in performance between the average student in LAC and their OECD counterparts equates to approximately five years of learning.

Our study seeks to evaluate an intervention that provided targeted remedial training in mathematics to 9th grade students, addressing a key stage in their educational progression. At this point, students are likely to have accumulated significant learning gaps that could affect their chances of completing secondary education. By focusing on this age group, the intervention aims not only to enhance math

[6] The 2023-24 school year officially started on August 28th 2023 and ended on June 21st 2024.

proficiency but also to promote retention and grade progression. This work aims to contribute to broader efforts aimed at improving learning outcomes and reducing dropout rates in the Dominican Republic and across the wider LAC region.

1.2.2 The Program: APRENDE +

The intervention studied, APRENDE + ("Acompañamiento Personalizado para el Rendimiento Escolar" or "Personalized Support for School Performance"), was implemented between October 2023 and May of 2024 in a sample of 38 9th grade classrooms. Figure 1.1 shows a full timeline of the study and intervention. The program had two core elements: a CAL intervention and a tutoring component. The first component introduced the use of a CAL platform (ALEKS) for two hours per week as part of schools' regular schedule, substituting two out of seven weekly hours scheduled for math instruction. Half of the classrooms were randomly selected to receive CAL licenses for all students and their math teachers (stratified by classroom size). The second component introduced small-group tutoring during extracurricular hours for approximately 20 percent of students in randomly selected classrooms from the first treatment group. Each component is summarized in more detail below.

By testing the intervention in schools, as part of the regular school day, we examine its potential for scale and assess whether this type of approach can enhance regular math instruction. The study and intervention was funded by the Tinker Foundation and was implemented in partnership with the Ministry of Education of the Dominican Republic (MINERD) and the Abdul Latif Jameel Poverty Action Lab (J-PAL).[7]

ALEKS platform. The program provided access to ALEKS, a math-focused CAL software developed by McGraw Hill. ALEKS uses predictive algorithms to deliver personalized learning experiences through adaptive assessments and assignments. These tools evaluate students' knowl-

[7] The study is registered as AEARCTR-0012383 in the AER registry, and was approved under protocol number 24-444 by the Teachers College IRB.

edge by identifying the math topics they master, the ones they do not, and the ones they are ready to learn. This approach is grounded in Knowledge Space Theory (KST), which posits that a subject's knowledge can be mapped as a set of solvable problems (Doignon and Falmagne, 1985; Falmagne et al., 1990). Accordingly, ALEKS assesses students' mastery by their ability to solve representative problems from each topic.

The software encompasses over 600 math topics for 9th graders. Because it would not be feasible to cover all topics within an academic year, this list was narrowed to topics that were aligned to the targeted grade (9th grade). This included topics outlined in the 9th grade curriculum as well as others that were considered prerequisites for this grade. This process gave us a total of 337 topics which were organized into nine modules, each corresponding to different content areas scheduled for instruction during the academic year, as per curriculum guidelines.[8] In addition to the practice exercises included in each module, ALEKS offers functionalities that enable teachers to assign and evaluate homework and tests. It also provides a dashboard through which teachers can track the progress of their classes and individual students.

All teachers assigned to a CAL classroom received an initial 2-hour training on the use of the learning technology. This training was provided by McGraw Hill, the developer of the learning platform. In addition to this, during the first 3 months of the program, teachers met one-on-one with a CAL aide for approximately one hour every two weeks. The CAL aide's role was to support them in accessing and monitoring students' assessments, tracking their progress, and identifying topics that students had not yet mastered for lesson planning. Importantly, one of the primary aspects of this extended training was to help teachers utilize the software's specific tools to ease their workload—including features that allow assigning and automatically grading homework—thus enabling teachers to save time that could potentially be devoted to other teaching activities.

Tutoring Intervention. The tutoring component aimed to explore complementarities between

[8]This selection was performed by two local experts on math pedagogy based on a revision of the curriculum. The majority of topics included as prerequisites were identified by the learning software.

CAL and more traditional forms of instruction, such as tutoring, by providing up to three weekly hours of tutoring to high-need students. Tutoring sessions were managed by one or two tutors, depending on the number of tutees per classroom, which typically ranged from 5 to 8. In each classroom selected to benefit from the tutoring component, we targeted approximately 20 percent of students (up to 8 students per classroom). At baseline, in order to select which students would receive tutoring, math teachers were instructed to select and rank up to 10 students for group tutoring based on need. To minimize bias in teachers' selection, after their first selection, we informed them of students' results in baseline math tests at the time of ranking. About 70 percent of teachers modified their selection after receiving this information.

Tutors were selected through a competitive process that included interviews conducted by the research team. All tutors were either recent graduates or in the final stages of earning a Bachelor's degree in Mathematics with a focus on secondary education, and had experience in using information technology for teaching. Tutoring sessions were integrated into students' schedules as elective workshops, covering approximately three academic hours per week, typically divided in one to three weekly sessions.[9] Elective workshops are offered for all students attending extended-day schools in the Dominican Republic. These types of schools offer school days that start at 8am and end at 4pm, and the workshops typically cover areas outside of the core curriculum, such as computing or art lessons.

To ensure quality and consistency in the tutoring sessions, each tutor was supported by a supervisor, who also had training and experience in mathematics education for high school students. The supervisor's role included reviewing individual session plans before each session, ensuring that the topics and instructional methods had been prepared before the session, and providing feedback to tutors on their session plans if needed. In addition, tutors maintained close communication with the mathematics teacher to align their sessions with the classroom curriculum and were instructed

[9] Out of 9 tutoring classrooms, 2 sections had one weekly meeting of 3 hours, 5 sections had two weekly meetings of 2 and then 1 hour, and 2 sections had three meetings for 1 hour each day.

to use the CAL platform for the tutoring sessions.[10] During the sessions, tutors reviewed topics covered in class during that week or revisited previous material, including pre-requisite topics that students found challenging. For instance, if a class was working on linear equations, a tutor may have reviewed the rules of signs before addressing the current topic.

The effectiveness of the tutoring sessions was monitored through unannounced visits by a member of the research team or the supervisor. These visits were intended to identify potential issues or areas for improvement. One-on-one meetings between the supervisor and the observed tutor were arranged after these visits to exchange feedback. Additionally, to monitor attendance, tutors were asked to complete a weekly survey to report the number of tutoring hours conducted for each class and student.

1.3 Implementation

1.3.1 Sample and Randomization

The intervention was administered in 9 public high schools of the Dominican Republic. Schools were selected based on whether they met the following conditions: (1) having an active computer lab, (2) having regular access to electricity and internet, (3) having an extended-day format,[11] and (4) having at least three 9th grade classrooms. In addition to this, we limited our sample to schools who had participated in *Programate*, a joint collaboration between MINERD and the World Bank that provided access to the ALEKS software to approximately 60 schools between 2018 and 2023. This set of conditions gave us an initial list of 34 schools. Between April and August 2023, the research team, with support of MINERD, contacted this initial list of schools via phone to confirm whether they met these conditions and to assess their interest in participating in this intervention.

[10]For example, tutors were encouraged to use the platform to identify topics where students were lagging behind and review these during tutoring sessions, or use the platform's accompanying exercises as practice. If a given session had a higher tutee/tutor ratio due to a tutor absence, the tutor was encouraged to alternate the use of CAL with small-group tutoring.

[11]Approximately 67 percent of public high schools in the Dominican Republic offered an extended school day format during the 2021-2022 school year (MINERD, 2023), where the school day starts at 8am and ends at 4pm.

Participation in the intervention was voluntary and all contacted schools expressed strong interest in participating. However, out of this initial list of 34 schools, only 14 confirmed having access to internet and a computer lab.

Between September 4th and 15th, a member of the research team conducted school visits to validate the information collected during the Spring and Summer terms. Only 9 of the 14 visited schools had access to internet and functional computers. All participating schools were located in Santo Domingo and Santiago. Being the two largest cities in the country, they also have the largest concentration of students.[12]

Treatment was assigned using a two-stage randomization process at the classroom level. Initially, classrooms were ranked by size and paired with the closest matching classroom in terms of size. Within each pair, one classroom was randomly assigned to the treatment group and the other to the control group, ensuring similarly sized classrooms across groups. This pairing and randomization process was repeated within the treatment group to assign half of classrooms to the group tutoring component. The study ultimately comprised 38 classrooms, with 19 assigned to receive the CAL software (treatment group) and the remaining 19 serving as controls.[13] Of the classrooms in the treatment group, 9 were further selected to receive additional group tutoring. The randomization design is summarized in Figure 1.2 and power calculations are included in Appendix A.2.1.

1.3.2 Implementation: CAL Component

The CAL intervention started on October 23rd and concluded on May 3rd.[14] During the first week of implementation, all treated students took a mandatory "Initial Knowledge Check" through the

[12]According to MINERD's Educational Statistics Yearbook for 2021-2022, approximately 45 percent of students were located in Santo Domingo and Santiago.

[13]Students in control classrooms received temporary licenses to access the software approximately 2 months after the endline data collection.

[14]This is equivalent to roughly 22 weeks, excluding winter, Holy Week, and two weeks where the use of the platform was interrupted due to training and data collection.

learning platform (this is the first activity in the platform). Based on the information from this initial assessment, the platform predicts the list of topics students master, do no yet master, and are ready to learn, and creates an individualized learning path for each student.

The implementation of the CAL component faced several challenges that affected the delivery of some sessions. The most prominent issues included unstable internet connectivity, insufficient availability of computers, frequent power outages, and unplanned school activities that interfered with the scheduled sessions.[15] To mitigate the loss of sessions caused by connectivity problems, wireless modems were provided to supervisors. Additionally, some schools struggled to provide enough devices for all students during the CAL sessions. To address this shortfall, tablets were allocated to supervisors for use in these schools. Another challenge was the frequent occurrence of school events and social activities, which were often unplanned or not accounted for in the school calendar. These events sometimes overlapped with the scheduled CAL sessions, leading to the cancellation or rescheduling of sessions and further disrupting the learning schedule.

1.3.3 Implementation: Tutoring Component

The first tutoring session took place on October 23rd and the last on April 26th. Tutoring sessions took place one to three days per week during extracurricular hours. In all but one school, tutoring sessions took place during hours allocated for optional workshops.[16] Optional workshops were meant to cover extracurricular content and varied in scope. Some examples of the type of courses covered during these hours were computing, music or dance lessons. The workshops and tutoring sessions would typically take place between 2 and 3:30pm. In total, 61 students were selected to participate in these tutoring sessions.

The implementation of the tutoring component also faced several challenges. School activities

[15]These issues are common across schools in the DR due to inadequate infrastructure.

[16]One of the nine schools did not offer optional workshops, so students stayed after school hours to receive tutoring sessions.

frequently disrupted the tutoring schedule, and unlike the CAL sessions, these missed tutoring sessions could not be rescheduled.[17] Additionally, classroom availability posed a significant issue, resulting in tutoring sessions being held in the school cafeteria or other open public spaces where noise and distractions were prevalent. Another challenge tutors encountered was sustaining student interest and participation. Some students were reluctant to join the tutoring sessions, leading to delays in starting the activities.[18]

1.4 Data

To evaluate the effectiveness of the program, we assess several key outcomes: students' academic performance, socioemotional skills, psychological well-being, and time use both inside and outside of school. Additionally, to better understand the potential drivers behind any observed effects, we examine changes in teachers' resource allocation based on student-reported information. These outcomes were collected using a combination of survey data, administrative records, data from the adaptive software, and brief math assessments. Each of these outcomes is described in detail below.

Student surveys and math assessments were administered at baseline, midline, and endline by the research team. Baseline surveys were collected in September 2023, followed by midline surveys in December and endline surveys in May 2024. Administrative records were collected and digitized after the end of the school year, between June and September 2024.

Academic performance. Our main outcome of interest is students' academic performance, as we want to examine whether the platform or tutoring component were effective in improving student learning in math. To do this, the research team administered math assessments at baseline, midline, and endline. These assessments were designed using items of varying difficulty levels, adapted from MINERD's national assessments. Endline tests were designed by a pedagogical expert and

[17] If a class missed a CAL session, teachers often reallocated other math hours that week to make up for the missed CAL session.

[18] If a large number of students were absent from the sessions but present at school, the tutor often spent several minutes at the beginning of the session trying to locate them.

followed the content and cognitive areas used on the 2019 Trends in International Mathematics and Science Study (TIMMS) 8th grade math assessments, which included seven content domains and three cognitive areas (Mullis and Martin, 2017).[19] We examine student performance in each of these content areas and cognitive domains, both at midline and endline. Moreover, we classify items according to their difficulty levels to assess differences in impacts by the grade level of the content.

In addition to the tests, between June and September 2024 we collected and digitized schools' administrative records containing data on students' end-of-year school grades for math and Spanish (to test for adverse effects), their attendance records, whether they dropped out during the school year, and whether they passed or failed the grade.

Socioemotional skills. As students increasingly rely on digital devices for learning and have fewer face-to-face interactions with their teachers, there is a risk that socioemotional or non-cognitive skills may be negatively impacted. In baseline and endline surveys we measured the following three skills: perseverance, locus of control, and grit. Perseverance is measured through a real effort task adapted from Alan et al. (2019). In this task, students first answered a logic question, and then chose one of three options for their next step: attempting another question of the same difficulty, tackling a more difficult question, or giving up. Their choice is used as our measure of perseverance. To measure locus of control, or the extent to which students believe they have control over the outcome of events in their lives, we used an adaptation of Rotter's locus of control scale (Rotter, 1966). Lastly, we measured grit using a 4-item scale adapted from Carlana and La Ferrara (2021). We build indices for the grit and locus of control measures such that the minimum and maximum values are 0 and 1.

Student well-being. The nature of the learning platform, in which each student works inde-

[19]The 8th grade TIMMS math assessments evaluated the following content domains: (1) Numbers, (2) Algebra, (3) Geometry, and (4) Data and Probability. The cognitive domains evaluated were: (1) Knowing, (2) Applying, and (3) Reasoning. TIMMS is only administered in the 4th and 8th grades.

pendently, may have led to less opportunities for teamwork or support than regular instruction, potentially making students feel more isolated in class. Moreover, increased screen time has been linked to lower psychological well-being among teenagers (Stiglic and Viner, 2019; Twenge and Campbell, 2018). We used questions from the Children's Depression Screener (ChilD-S) developed by Frühe et al. (2012) to measure students' well-being.

Furthermore, by increasing students' math skills, or their awareness of their math skills, access to CAL may have mixed impacts on math anxiety. Following Araya et al. (2019), we created a math anxiety index using students' answers (on a Likert scale) to the following statements: 1) I get nervous before math tests and 2) I am worried of having bad grades in math.

Teacher behavior. The CAL platform may free up teachers' time by automating some of their daily tasks, such as developing and grading homework and tests. As a consequence, we may observe differences in the allocation of teachers' time and effort across classroom activities. To assess changes in these outcomes, we ask students about their perceptions of teacher behavior in the classroom (e.g. did the teacher give you additional support outside of the classroom?, did they clarify a lesson that you did not understand?) and their time management (e.g. did the teacher start class late?, did the teacher come to school?). We use their answers to build four indices of teacher behavior: 1) time mismanagement index, 2) teacher engagement index, 3) teacher effectiveness index, and 4) teacher ineffectiveness index. The first index, on time mismanagement, is based on students' answers to a series of questions on teachers' timeliness to class and school. The third index is built upon students' answers to a set of questions around their engagement during class.[20] The last two are based on classroom activities, which we classify into "effective" and "ineffective" practices, depending on their correlation with control group students' midline and endline test scores. Practices that are positively correlated with test scores are combined into an "effective

[20] The questions used to build this index are: 1) Your math teacher keeps students engaged, 2) Asks questions to check for understanding, 3) Knows when everyone understands, 4) Makes every student work hard, 5) Assigns homework that helps you learn, 6) Reviews past topics, 7) Marks your work to help understanding, and 8) Wants you to explain your answers.

practices" index and those negatively correlated are used for an "ineffective practices" index.[21] All indices are standardized to have a mean zero and a standard deviation of 1 in the control group. The questions used have been adapted from questionnaires used by De Hoyos et al. (2021) and Ferguson and Danielson (2015).

Software data is used for both monitoring and assessment purposes. The software contains information on students' interactions with the software, including the number of sessions held, the date and duration of each session, and the number of newly mastered topics by each student. We use this information to assess compliance with the intervention and to measure progress within the treatment group. In addition to this, we use the software data to measure the frequency with which teachers use it to assign and grade homework, in-classroom activities, and assessments. This data is available for the treatment group only.

1.4.1 Balance Checks and Attrition

We present balance checks in Table 1.1 using information from baseline surveys. There are no statistically significant differences between the treatment and control groups across observable characteristics. Approximately 56 percent of the sample is female. The average age of students in our sample is close to 15 (the typical age for a students enrolled in 9th grade is 13 to 14 years old), reflecting the high levels of grade repetition in the country (approximately 19 percent of students in our sample repeated at least one grade). Baseline scores are relatively low: students answered an average of 6 out of 12 questions correctly. Moreover, we can observe that the average student in our sample comes from a low socioeconomic background; for instance, less than 40 percent of the sample has a mother or father who attended college.

As shown in Columns 1 and 5 of Table 1.2, survey and test completion rates were high at both

[21]The "effective" practices are: 1) You used a textbook, 2) You copied from the board or book, 3) You solved problems, and 4) You practiced problems on the board; while those classified as "ineffective" are: 1) You worked in a group, 2) Your teacher explained math topics, and 3) Your teacher asked you to practice for a test.

midline and endline: 92 percent of the baseline sample completed surveys and tests at midline, and 90 percent did so at endline. The differences in completion rates between the control and treatment groups are small: 0.5 percentage points at midline and 2.9 percentage points at endline. While the latter difference is statistically significant at the 5 percent level, it is small enough not to raise substantial concerns.

Most importantly, the students who completed the surveys appear similar to those who did not. Columns 2-4 and 6-8 show minor differences in observable characteristics between respondents and non-respondents. For instance, treatment group students who responded to endline surveys were, on average, 0.037 years younger (less than half a month) than non-respondents, and the proportion of female respondents was 4 percentage points lower. Baseline test scores were 0.01 standard deviations higher among respondents in the treatment group. Overall, we believe the low rates of attrition and minimal differences between respondents and non-respondents do not compromise the integrity of our findings. In Appendix Table A.4, we present Lee (2009) bounds for the main performance outcomes. The results indicate that selection due to differential attrition does not account for the observed effects, as these appear similarly large after applying this bounding method, with gains of 0.09-0.32 standard deviations at midline (not statistically significant) and 0.23-0.37 at endline for the CAL-only treatment. The tutoring interaction remains statistically insignificant.[22]

1.5 Empirical Framework

We estimate the impact of the program on student outcomes using the following regression model:

$$Y_{ci} = \beta_0 + \beta_1 \text{CAL}_c + \beta_2 \text{Tutoring}_c + \beta_3 X_{ci} + \epsilon_{ci}$$

[22] We calculate the lower bound by dropping the fraction of highest scoring test-takers in each treatment group until the response rates are equal across the three experimental groups. To estimate the upper bound, we drop instead the fraction of lowest-scoring students. We do the opposite when response rates are highest in the control group.

where Y_{ci} is the outcome of student i in classroom c and CAL_c is a binary variable that takes the value of 1 if classroom c was assigned to treatment and 0 otherwise. $Tutoring_c$ takes the value of 1 if the classroom was assigned to the group tutoring treatment and 0 otherwise. X_{ci} is a vector of baseline characteristics for student i in classroom c. In this specification, β_1 captures the average treatment effect (ATE) of the CAL software, while β_2 captures the additional impact ("value-added") of the tutoring component relative to classrooms that received only the CAL software. In the results section we present each coefficient separately and the sum of the two. The latter represents the effect for classrooms assigned to both the CAL and tutoring interventions, relative to the control group.

Throughout this paper, we use randomization-based inference as our preferred methodology for analysis, which has gained prominence over traditional sampling-based methods in recent years (Athey and Imbens, 2017; Green et al., 2011; Young, 2019). This approach leverages the experimental design to derive statistical inferences directly from the randomization procedure, ensuring that our methods align closely with the study's design. Randomization inference is particularly advantageous in the context of nonstandard randomization techniques, such as the pairwise randomization used here, as it allows us to account for the specific structure of the randomization. Moreover, it enables us to obtain nearly exact p-values without relying on assumptions about the shape of the sampling distribution, which is especially helpful in small samples or when the number of clusters is limited, as is the case in this study. In such situations, traditional assumptions about the sampling distribution may be less reliable, making randomization inference a more robust approach.

To test the sharp null hypothesis of no treatment effect, we use a potential outcomes framework. In this framework, the potential outcome for each individual i in classroom c depends on two factors: the classroom's assignment to the CAL software (Z_c) and the classroom's assignment to the tutoring intervention (W_c):

- Z_c is a binary variable indicating whether classroom c was assigned to receive the CAL software ($Z_c = 1$) or not ($Z_c = 0$).
- W_c is a binary variable indicating whether classroom c was assigned to the tutoring intervention ($W_c = 1$) or not ($W_c = 0$). Note that within classrooms assigned to the tutoring intervention, only 20 percent of students received the actual tutoring, but the assignment W_c was made at the classroom level.

The potential outcomes are defined as follows:

- $Y_{ci}(1,1)$ represents the outcome for individual i in classroom c if the classroom is assigned to receive both the CAL software ($Z_c = 1$) and the tutoring intervention ($W_c = 1$).
- $Y_{ci}(1,0)$ represents the outcome for individual i in classroom c if the classroom is assigned to receive the CAL software ($Z_c = 1$) but not the tutoring intervention ($W_c = 0$).
- $Y_{ci}(0)$ represents the outcome for individual i in classroom c if the classroom is assigned to the control group ($Z_c = 0$), with neither the CAL software nor tutoring.

The sharp null hypothesis posits that these potential outcomes are identical regardless of the treatment condition, meaning the treatment has no effect. Formally, this can be stated as:

$$Y_{ci}(1,1) = Y_{ci}(1,0) = Y_{ci}(0) \quad \text{for all individuals } i \text{ in all classrooms } c$$

This implies that any observed differences in outcomes between the groups are due solely to the random assignment process and not to the intervention itself. By simulating 2000 randomizations, we approximate the sampling distribution of the ATE under the sharp null and estimate the probability of observing an ATE at least as large as the one found in the actual experiment if the sharp null were true. This simulation process is repeated for each outcome analyzed.

The validity of our findings relies on the "no-interference" or Stable Unit Treatment Value Assumption (Rubin, 1978), which assumes that the treatment of one classroom does not affect the potential outcomes of students in another. The platform used in this study required each student to log in using a unique username and password, and its use was closely monitored during the designated period for this task. These measures minimize the likelihood that students in non-treated classrooms could have accessed or benefited from the platform, thereby supporting the assumption of no interference. However, it is conceivable that treated students may have interacted with control students in ways that influenced their outcomes (e.g., through assistance with homework or exam preparation). Additionally, the assignment of a classroom to treatment may have impacted how teachers interacted with students in non-treated classrooms. We do not have anecdotal evidence suggesting that teachers changed how they interacted with control group classrooms as a result of the treatment, or that treated students collaborated with peers from different classrooms (based on conversations with the teaching team).

1.6 Results

1.6.1 Tutoring and CAL Use

The results from the initial CAL assessment suggest that, at the time of completion, the median student mastered only 2.7 percent (9) of the 337 math topics included as part of the course. Throughout the school year, the median treated student spent approximately 14 hours (approx. 38 minutes per week) on the platform and learned an average of 45 new math topics through it. As shown in Figures 1.4 and 1.5, the distribution of the number of hours and the number of topics learned on the platform is left-skewed, with the top 10th percentile of students spending over 38 hours using the platform and learning more than 188 new math topics throughout the school year.

To examine whether students used the CAL platform outside of school, we assess the share of time where the student used the software outside of the scheduled days. Note that this is a conservative estimate, given that teachers had flexibility to reschedule the use of CAL if classes had been cancelled during the scheduled days. Using this proxy, we find that for the median student 15 percent of the time spent using the platform took place outside the regularly scheduled days. This is equivalent to approximately two hours total. Given these numbers, we safely conclude that any effects observed come from a substitution of the time spent during class and not from additional instruction time.

Of the 61 students assigned to the tutoring treatment arm, all received at least 2 hours of tutoring, with the median student receiving 29 hours, or approximately 1.4 academic hours per week. Note that this is significantly below the intended dosage of 3 hours per week. This is partially explained by student absenteeism during the tutoring sessions. On a given week, up to 47 percent of students failed to show up to at least one of the meetings, based on monitoring data. We include a more detailed discussion of this and other implementation challenges in section 1.7.

1.6.2 Program Effects

Student Performance

We present Intention-to-Treat (ITT) effects roughly two (midline) and six months (endline) after the start of the program in Table 1.3. Columns 2 and 3 present results at midline and columns 4 and 5 at endline. We find that by midline the CAL-only treatment had produced gains of approximately 0.17-0.21 standard deviations, as shown in Columns 2 and 3. Although large, these results are not statistically significant. These impacts are roughly consistent across specifications (with or without covariates). We do not find significant effects from the tutoring treatment at midline, likely because students had only received 8 tutoring sessions at this stage.

By endline, CAL-only effects had increased to 0.29-0.31 standard deviations (statistically significant at the 5 percent level), as shown in Columns 4 and 5. The coefficient associated with tutoring was negative and relatively large at -0.19 when including covariates. Although not statistically significant, the magnitude of the coefficient is a cause for concern and we explore in the following section some hypotheses behind these results. We believe this negative coefficient may partly reflect disruptions to the tutoring sessions due to scheduling conflicts with other school activities and a lack of student engagement (including high levels of absenteeism). Additionally, tutor availability issues resulting from weather and transportation challenges likely further impacted the consistency and quality of tutoring sessions. Effects on performance from the CAL component align with those found in other work using computer-assisted instruction in low or middle income countries.[23]

When looking at school grades by the end of the year, we find small negative differences on students' math grades for CAL-treated students, relative to the control group; however, these differences are not statistically significant (see Table 1.8). We present effects on Spanish as a robustness check for the previous result. Interestingly, we do not observe similar differences for Spanish grades. The observed patterns hold whether we condition grades on having finished the school year (Panel B) or use the average of available periods for students who dropped out and/or are missing information for at least one school period (Panel A).[24]

Intrinsic Motivation, Absenteeism, and Dropout

Beyond academic performance, improvements in learning could also positively influence students' retention in school by increasing their motivation and interest in the subject, as well as their chances of obtaining a passing grade. In Tables 1.4-1.7, we present impacts on intrinsic motivation, preferences towards math, absenteeism, and student dropout. Student dropout is a significant concern in our context, especially among the targeted age group. In the Dominican Republic, over 40 percent of students drop out between the 8th and 12th grades (IDEICE, 2017). In the control group, 8 percent of students dropped out of school during the intervention period (see Table 1.7).

[23]For example, Araya et al. (2019) report effects of 0.21-0.27 standard deviations (SD) from a computer-assisted, game-based math learning platform in Chile after seven months of program exposure. In India, Muralidharan et al. (2019) find effects of 0.37 SD after 4.5 months. Finally, Büchel et al. (2022) observe effects of up to 0.17 SD from a computer-assisted learning software in El Salvador over a six-month period.

[24]The school year is divided in four quarters. Final grades are estimated as the average of each quarterly grade.

However, as shown in Tables 1.4-1.7, we find no evidence that the intervention impacted students' intrinsic motivation, how much they like math (self-reported 1 to 4 scale), school absenteeism, or dropout rates by endline. The lack of findings on intrinsic motivation align with Araya et al. (2019), who also observed no effect on this outcome in a study using a game-based learning platform in Chile.

Math Beliefs and Self-Concept

As students spend more time on the software and master new topics, they receive new information on their abilities and may update their beliefs accordingly. In Tables 1.9 and 1.10, we present effects on students' beliefs about their math performance (expected test scores), their confidence in the subject, and their math self-concept. Panel A of Table 1.9 shows that CAL students expected higher test scores than the control group (by 0.12-0.13 SD). However, this gap is not statistically significant and is not as large as the one observed in actual test scores (0.3 SD), suggesting that this group may have downwardly biased beliefs about their endline performance.

We do not find differences in expected performance between CAL-only classrooms and those that received tutoring, but Appendix Table A.1 shows that students in the latter group were more likely to be overconfident in their performance, although these results are not statistically significant.[25] Despite the lack of statistical significance, the differences in magnitudes suggest a potential for overconfidence in the CAL+Tutoring group and underconfidence in the CAL-only treatment. Panel C in Figure 1.3 illustrates the gap between students' expected performance and their actual performance on endline math tests, expressed as a share of correct answers. The largest gap is observed in the CAL+Tutoring group, followed by tutees and the control group. The smallest gap observed corresponds to the CAL-only group. These findings mirror those of Table 1.9 (Panel B), which shows that the CAL-only and CAL+Tutoring groups had similar scores on a math self-confidence

[25] The overconfidence outcome takes the value of 1 if students' estimated performance on the math test exceeds their actual performance.

index, built upon a series of questions on how comfortable they feel learning and applying math.[26]

Lastly, we examine math self-concept, defined as students' perceptions of their own ability in mathematics compared to others (see Table 1.10). As in Araya et al. (2019), despite the large increases observed in math performance, we do not find statistically significant differences across experimental groups in a math self-concept index or on individual measures used to build the index. However, students in the CAL+Tutoring group were 9.3 percentage points less likely to agree with the statement "Math is harder for me than for many of my classmates" than students in the control group or in the CAL-only group. While these results are not conclusive, they may be suggestive of shifts in students' self-perception and confidence in their math abilities. Despite performing worse on math tests, the tutoring group seems to feel equally confident about their performance and are more likely to have mismatched expectations about their outcomes in the test, when compared to those whose classrooms were not assigned to the tutoring component.

Self-Efficacy, Grit, Perseverance, and Locus of Control

By offering personalized learning pathways and continuous feedback on individual progress and milestones, the platform allows students to understand and learn from their mistakes privately, without the stigma of public correction. As students master new topics and receive real-time feedback, this process may enhance their sense of competence and self-efficacy. In this subsection, we examine four related outcomes: self-efficacy, grit, perseverance, and locus of control. As shown in Table 1.12, neither the CAL intervention nor the tutoring intervention had a significant impact on a self-efficacy index.[27] This suggests that the interventions did not notably alter students' beliefs

[26]The outcome variable for the self-confidence table is a standardized index based on students' answers to the following items: 1) When teachers teach me a new math topic, I learn it without much difficulty; 2) When I take a math test, I am confident that I can answer the questions well; 3) I can help my classmates with math homework that they don't understand; and 4) I feel more capable as I learn more things in math.

[27]This index is built using students' answers to the following items: 1) I can always solve difficult problems if I try hard enough, 2) I can solve most problems if I invest the necessary effort, 3) I can stay calm when I face difficulties because I can trust my coping skills, 4) When faced with a problem, I can usually find several solutions, and 5) If I'm in trouble I can find a solution.

in their ability to achieve academic success.

Similarly, we do not find impacts on a measure of grit (see Table 1.13), which assesses students' perseverance and passion for long-term goals (Duckworth et al., 2007). Following Alan et al. (2019), we implement an alternative measure of perseverance, based on a real-effort exercise. Table 1.14 shows that treated students are 8.7 percentage points more likely to choose to undertake an easier task against a more challenging alternative. We do not observe differences in their likelihood of giving up after failure for either group. This poses the possibility that the adaptive nature of the platform, which presents exercises based on students' current ability, may have negatively impacted their willingness to take on more challenging exercises.

Lastly, we examine effects on locus of control, which refers to the degree to which individuals believe they have control over the outcome of events in their lives. A higher internal locus of control indicates that individuals feel they can influence their own success, while an external locus of control suggests that they attribute outcomes to external forces. Table 1.15 shows that the CAL intervention had no significant impact on students' locus of control. However, the tutoring component had a small but statistically significant negative effect on this outcome (0.02–0.03 percentage points), indicating that students in the tutoring group were slightly more likely to feel that their performance was determined by external factors rather than their own actions. We include a discussion on this point in the following section.

Student Wellbeing

To assess changes on students' wellbeing, we use two proxies: the Children's Depression Screener (ChilD-S) and a a short questionnaire on math anxiety. For the first one, we build an index based on students' answers to 8 items about their mental state (e.g., "I feel happy", "I feel lonely", "It is hard for me to concentrate"). For the second one, we use two questions on students' worry about their math grades and their feelings before math tests to build a math anxiety index. Table **??** shows that the intervention had no impact on the first measure of wellbeing. While we do not find impacts on a math anxiety index, we find important differences when looking at individual measures of math anxiety. As seen in Table 1.18, students in the CAL-only group were 6 percentage points more likely to get nervous before math tests than students in the control and CAL+Tutoring groups. This difference is statistically significant at the 5 percent level. While math anxiety is typically linked to lower performance (Dowker et al., 2016), this result may be positive if it signals higher levels of awareness or concern about their math performance in this group.

In the Appendix, we present results on gender beliefs (Table A.3) and student time use (Table 1.16). We do not find statistically significant effects on these outcomes.

Teacher Practices and Time Use

In this section, we estimate the effects of the intervention on teachers' allocation of time and effort across various activities, as this is one potential mechanism by which the intervention may have impacted teacher effectiveness and student performance. In Panel A of Table 1.19, we show impacts on a time mismanagement index based on students' responses to a series of questions on teachers' timeliness to school/class.[28] We do not find statistically significant differences on this index across treatment groups. In Panel B we assess effects on an index of teacher engagement, built upon students' answers to questions around their engagement during class.[29] Although the index is lower for treated students, relative to the control group, these differences are not statistically significant.

Lastly, we look at classroom activities, which we classify into "effective" and "ineffective" practices, based on their correlation with student midline and endline test scores. Practices that are positively

[28]The questions used to construct this index are: 1) Was your teacher late for class?, 2) Did they not come to school?, 3) Did they start the class early?, 4) Did they finished class early?, 5) Did they finished the class late?, and 6) Did they left school early?

[29]The questions used to build this index are: 1) your math teacher keeps students engaged, 2) Asks questions to check for understanding, 3) Knows when everyone understands, 4) Makes every student work hard, 5) Assigns homework that helps you learn, 6) Reviews past topics, 7) Marks your work to help understanding, 8) Wants you to explain your answers.

correlated with test scores are combined into an "effective practices" index and those negatively correlated are used for an "ineffective practices" index. As shown in Table 1.20, we do not find statistically significant differences across experimental groups on these outcomes.

Overall, the lack of findings on teacher behavior suggest that the effects observed may be attributed to the substitution of some of the teachers' instructional time with the new technology (skill replacement), rather than a reallocation of teachers' time and effort (or that of their students) across different tasks. This is also consistent with our analysis of platform data. Although teachers spent an average of 33 minutes per week actively using the platform, they seldom leveraged it to automate routine tasks such as assigning or grading homework and exams. Our data indicates that, over the course of the academic year, teachers assigned no more than three activities through the CAL platform.

1.6.3 Heterogeneity in Program Effects

As students progressed through the software, they were targeted with exercises that catered to their learning levels. Following Muralidharan et al. (2019), we classify test topics into two categories–below and at grade level–and test whether students fare differently across these two categories in the tests. Table 1.21 suggests that the CAL treatment was effective in improving performance on both below and at grade level content, with impacts ranging from 0.21 to 0.28 standard deviations. This pattern differs from that found in Muralidharan et al. (2019), which finds large impacts on below grade math items but no impacts on at grade level items.[30] However, the patterns are consistent with what we would expect by looking at software use, as the software was designed to cover both below and at grade level content. At midline, when students had spent most time on modules covering below grade level topics, we find that effects were largely concentrated on test items covering this type of content.

[30]In Hindi, the authors find significant effects for questions at and below grade level.

In Table 1.22, we present effects across content areas included in TIMMS mathematics assessments for 8th grade. We observe positive differences from the CAL treatment on the algebra, numbers, and geometry content areas by endline, although the coefficient for geometry is not statistically significant. We do not find impacts on data and probability, the one TIMMS content area that was not covered in the platform or in the Dominican Republic's 9th grade curriculum. For the tutoring component, results seem to mirror those found on Table 1.3, as classrooms assigned to a tutor show negative but statistically insignificant differences, when compared to the CAL-only treatment, across all content areas except for data and probability. Beyond content areas, TIMSS also evaluates students' abilities in three cognitive domains: knowing, applying, and reasoning. In Table 1.23, we present results across these cognitive areas. As shown, we find large and positive effects on all cognitive areas, ranging from 0.16 to 0.31 standard deviations, although results are not statistically significant for the "knowing" cognitive area.

Lastly, Figure 1.6 presents the distribution of performance for both the treatment and control groups across baseline performance percentiles. Although confidence intervals overlap at several points, we observe that the treatment group consistently outperformed the control group at all levels of baseline performance.

1.7 Discussion

In this section, we discuss how the implementation challenges we faced when deploying CAL in schools could have affected the results observed on performance, and delve into a key remaining question: why was the tutoring component ineffective? While we do not have conclusive evidence to explain these findings, we present in this section some potential explanations.

1.7.1 CAL Implementation Challenges and Impacts on Take-up

Challenges mentioned in subsection 1.3.2 resulted in lower-than-expected use of the CAL platform, with the average student using the platform for less than 60 percent of the intended time. In Appendix Table A.5, we estimate the dose-response relationship between hours of CAL use and value-added (Muralidharan et al., 2019). We instrument for hours of use using the random assignment to the experimental treatment arms. The IV estimates indicate that, on average, one additional hour of CAL use increased endline test scores by 0.027 standard deviations. Figure A.2 shows that diminishing returns start to kick in around the 1500-minute mark. Up until that point, the slope of the curve increases steeply, but then the rate of increase in math test scores begins to taper off, signaling that additional time on CAL has progressively smaller benefits on test scores. Note that the effect identified here averages causal effects over varying levels of treatment. Thus, an extrapolation of up to 1500 minutes of use is likely to be downwardly biased, whereas an extrapolation of more than 1500 minutes of use may overstate the true effects of the intervention. In Table A.5 we limit our extrapolation to 1500 minutes of use, which, assuming 80 percent attendance, would be equivalent to roughly 5 months of implementation.[31] Our results suggest that implementing the program with fidelity for 5 months would lead to test score gains of 0.73 standard deviations.[32]

1.7.2 Ineffectiveness of Tutoring Component

The challenges mentioned in subsection 1.3.3 help explain the limited effectiveness of the tutoring sessions. By midline, the average attendance at tutoring sessions was approximately 75 percent. In an effort to address this, students were given the option early in the spring to withdraw from tutoring and join an alternative extracurricular workshop if they preferred. While the majority (54

[31] Most students (approx. 90 percent) fall below the 1500-minutes mark, which reduces our concern that diminishing returns may bias these estimates. Moreover, we cannot reject a test of equality (difference-in-Sargan) between the IV and OLS estimates using a value added specification, and as seen across columns 1-3, the coefficient associated with hours of use is similar across model specifications, suggesting that the LATE may be similar to the ATE.

[32] This extrapolation requires the assumption of heterogeneity in treatment effects. As shown in Figure 1.6, we do not find evidence of effect heterogeneity across baseline performance.

out of 61 students) remained enrolled in the tutoring program through the end of the school year, 7 students opted to withdraw at the beginning of the spring semester. Despite the relatively low dropout rate, attendance continued to pose a challenge, with average student attendance standing at 73 percent by the end of the intervention. Using a meta-analytic sample of 265 tutoring studies, Kraft et al. (2024) find that program features that change as tutoring programs are taken to scale, including higher tutor-tutee ratios and shorter delivery dosages, partially account for the reduced effectiveness of these programs (when compared to smaller scale interventions).[33]

Moreover, if the program disrupted other productive learning activities without delivering the consistent support needed to drive academic improvement, it could have had a net negative impact. While no observable differences were found between the tutoring group and the control group, we did observe negative (though not statistically significant) differences when compared to the CAL-only group, as well as a reduced usage of the CAL platform. This issue is explored in greater detail in the following subsection.

1. Use of CAL platform. We find that students in classrooms assigned to the tutoring arm spent less time on the platform and put lower effort while using it.[34] Figure 1.9 shows that the average student assigned to a CAL-only classroom spent a total of 1232 minutes using the platform (56 minutes per week) while the average student in a CAL+Tutoring classroom spent 1082 minutes (49 minutes per week). We use the relationship between minutes of use and test scores to predict how this difference in use may affect results. Assuming a linear dose-response relationship between the two, we find that this difference in time use may account for approximately 34 percent (0.07 SD) of the observed difference in endline performance between classes with tutoring and those that only included the CAL component (see Appendix Table A.5). Within CAL+Tutoring classrooms, both non-tutees and tutees had considerably lower levels of use than the average student in a CAL-only classroom, but the difference in use was significantly lower for tutees, who spent only 1000 minutes

[33] Kraft et al. (2024) show a declining pattern of effects as student-tutor ratios increase.

[34] We define effort as the ratio of topics learned to time spent on the platform.

(or 45 minutes per week) using the platform.

2. Selection of Tutees. Although we anticipated that teachers would select the lowest-performing students as tutees, the data show otherwise. The average tutee was not in the bottom 20th percentile on baseline tests; in fact, as Figure 1.7 illustrates, the average tutee was located in the second quintile of performance, approximately 0.7 standard deviations below the average. This discrepancy may stem from the timing of the selection process, which occurred at the start of the school year when teachers may have had limited knowledge of their students' abilities. A plausible explanation is that initial tutee selection was based on incomplete or inaccurate information, or that teachers considered students' likelihood of attendance as one selection criteria, perhaps assuming that the worst performing students were less likely to attend the tutoring sessions.[35] It remains unclear whether teacher judgment or baseline test scores should serve as the optimal criteria for selecting tutees. Nevertheless, as Figure 1.8 shows, students selected as tutees were closer to the mean than to the bottom, and exhibited flatter growth in test scores between the baseline and endline than other comparable groups, suggesting that the selection process may not have effectively targeted the students who would have benefited the most from the tutoring intervention.

3. Locus of Control. Students in the CAL+Tutoring group were less likely to score positively on the locus of control variable (Table 1.15). The significance of locus of control becomes apparent when we examine the effects by students' baseline levels on this variable. When the sample is split into high (above the baseline mean) and low (below the baseline mean) locus of control groups, Table 1.24 shows that the large positive effects observed in the CAL-only group are concentrated among students with initially high levels of locus of control, while the negative effects in the tutoring group are concentrated among those with lower initial levels on this variable. A possible explanation is

[35] After the initial selection, teachers were provided a list of the bottom 20 percent of students in their class, along with their baseline test scores. While 70 percent of teachers made at least one adjustment to their tutee selection, only 60 percent of the students from the original list were changed. Only 41 percent of the students on the updated lists were from the bottom 20 percent in terms of baseline performance (an increase of 20 percentile points compared to teachers' initial selection). The most frequently cited reason for selecting a tutee was having low performance (82 percent), followed by showing behavioral issues (6 percent) or lacking an appropriate study method (6 percent).

that the presence of a tutor may have led students to perceive them as responsible for their academic success, making them more confident about their outcomes and more passive in their efforts. This could explain the lower use of the CAL platform among students in tutoring classrooms, as they might have relied more heavily on the tutor's support. In Table 1.18, we show that students in the CAL-only group were more likely to report getting nervous before math tests than students in the control and CAL+Tutoring groups.

1.8 Conclusion

This paper contributes to the growing literature on computer-assisted learning by testing an innovative intervention that replaces nearly one-third of regular math instruction with the use of an adaptive learning platform. Conducted in the Dominican Republic, a lower-middle-income country, this study highlights both the challenges and transformative potential of introducing technology-based tools in public school classrooms. We find that substituting two weekly hours of regular instruction with CAL use significantly improves student performance on math tests by up to 0.3 standard deviations, with these gains being consistent across various difficulty levels and subject areas.

Interestingly, when the CAL component was combined with a tutoring intervention targeted at 20 percent of students in each classroom, the effectiveness of the CAL component was reduced, even for those who did not directly participate in the tutoring sessions during extracurricular hours. Implementation challenges, such as low attendance rates among tutees and reduced platform usage and effort among students in the tutoring group, may explain this diminished effectiveness. Although we do not have conclusive evidence for why platform use decreased, we find impacts on student locus of control and wellbeing, with students in the tutoring group exhibiting lower locus of control and lower math anxiety compared to those who only received the CAL component. While these results do not fully account for the reduced performance observed in the tutoring group, they underscore the need to better understand how technology-based interventions interact with other classroom supports.

Overall, this paper illustrates the promising potential of technology to enhance education systems worldwide. However, it also highlights the importance of understanding the dynamics between technology integration and concurrent interventions, such as tutoring, within the school context. Furthermore, we highlight the importance of infrastructure and logistical planning in the successful implementation of programs that take place in schools and rely on the use of technology, underscoring the need for continuous monitoring and adaptive strategies to minimize disruptions. As part of future work for this project, we plan to conduct qualitative research, including focus groups and interviews with teachers and students, to explore potential shifts in attitudes or behaviors resulting from the tutoring treatment arm.

Figure 1.1: Timeline for Intervention and Study

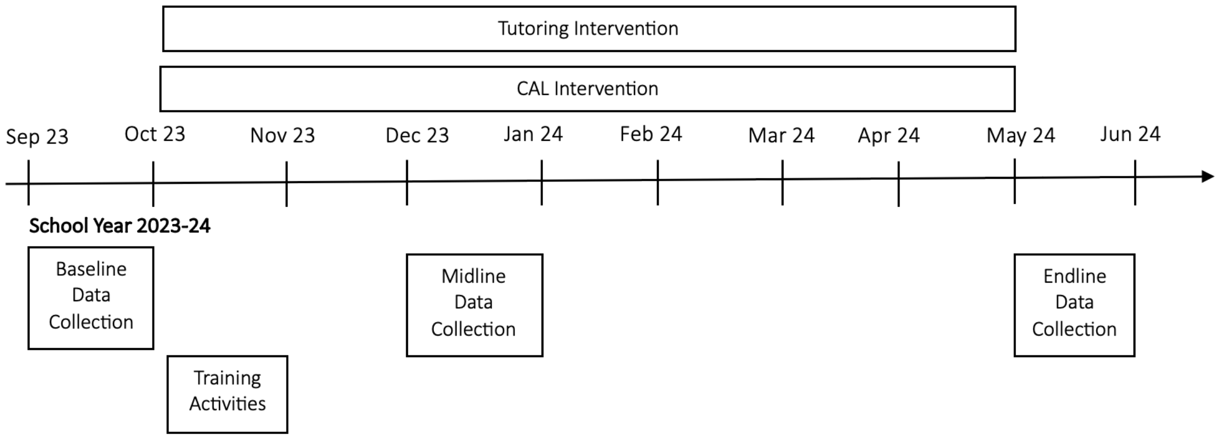
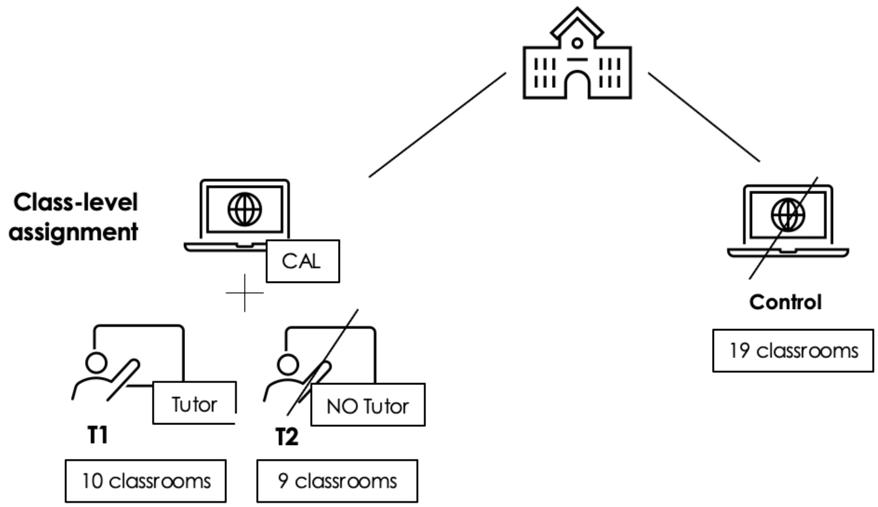
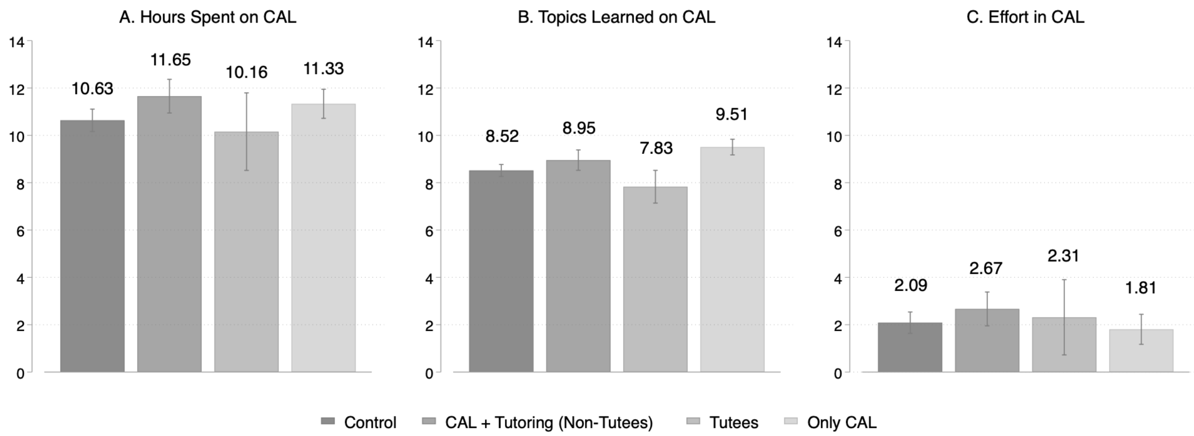


Figure 1.2: Randomization Design



Notes: T2 is composed of 9 classrooms assigned to receive access to the CAL software. T1 is composed of 10 classrooms where in addition to receiving access to the software, 20 percent of students, were to receive tutoring sessions in small groups during extracurricular hours. The control group did not receive access to the CAL or tutoring components.

Figure 1.3: Expected Performance on Endline Math Tests



Notes: The CAL+Tutoring group excludes tutees. Panel C shows the difference between students' reported expected performance (share of correct answers) and their actual performance.

Figure 1.4: Time Spent on CAL

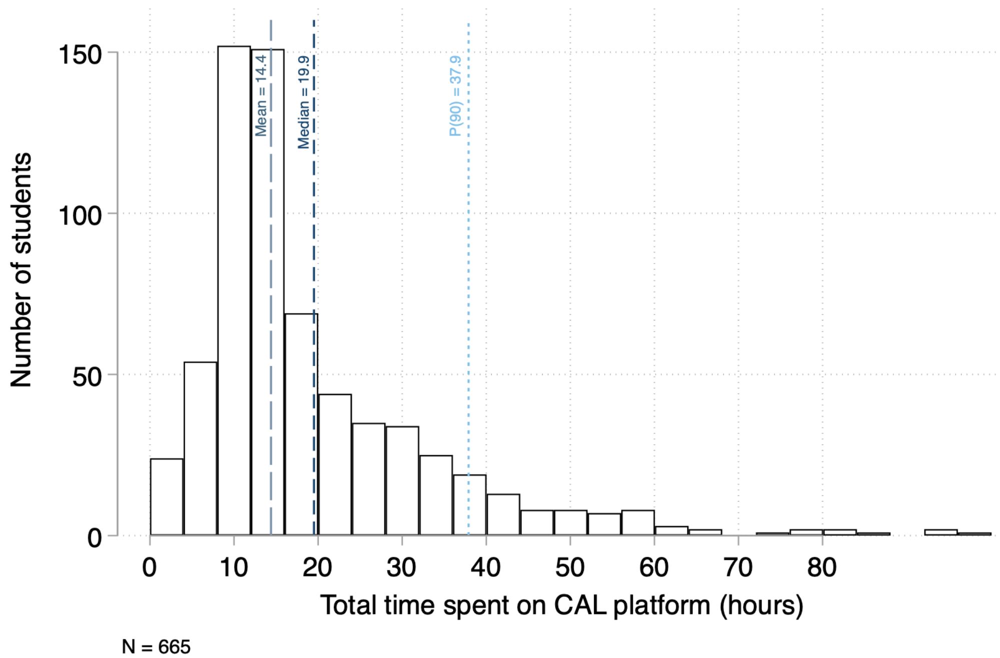


Figure 1.5: Topics Learned on CAL

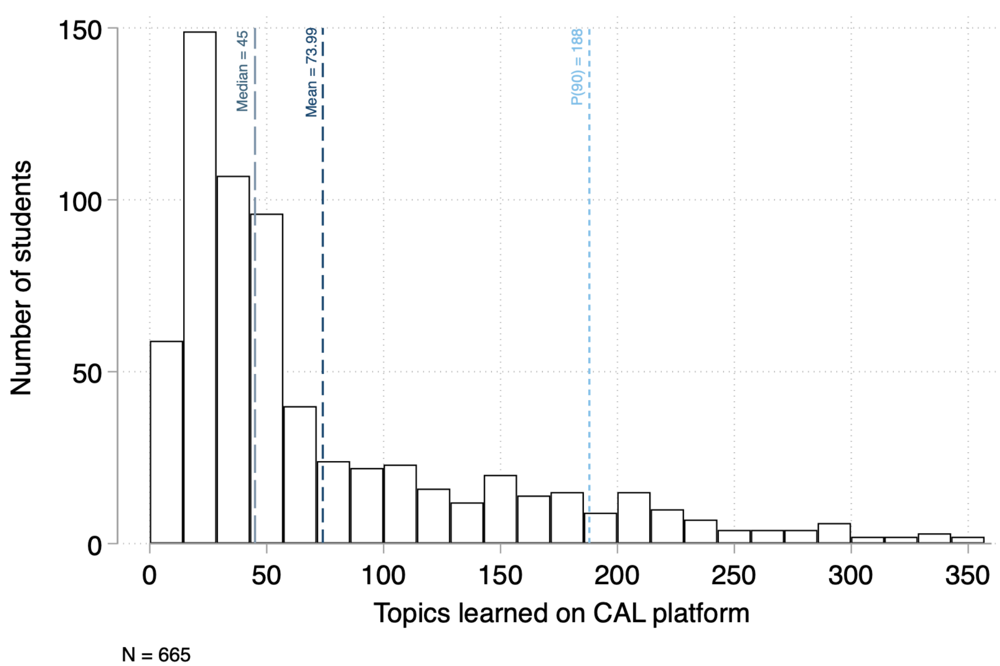
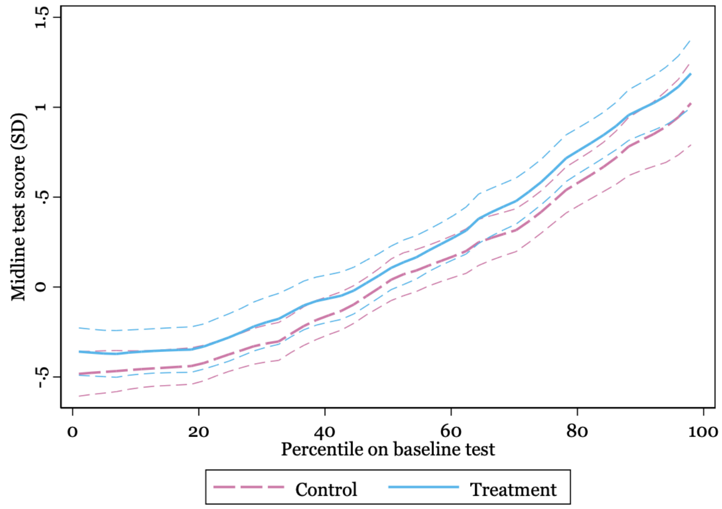
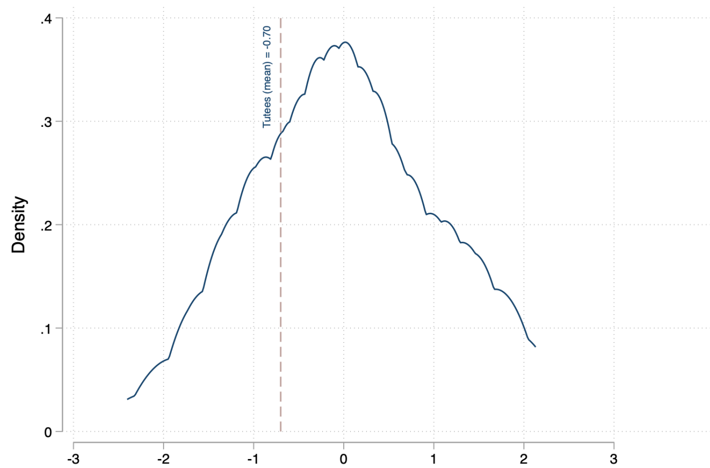


Figure 1.6: Non-Parametric Estimates of Treatment Effects by Baseline Performance



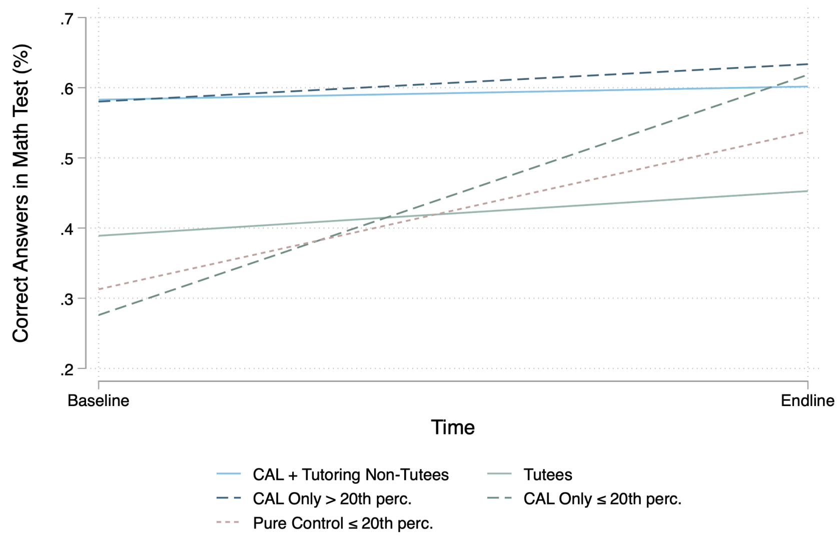
Notes: Following Muralidharan et al. (2019), the figure presents kernel-weighted local mean smoothed plots relating endline test scores to percentiles in baseline tests, separately for the treatment and control groups. The dashed lines show 95% confidence intervals.

Figure 1.7: Selection of Tutees and Distribution of Performance



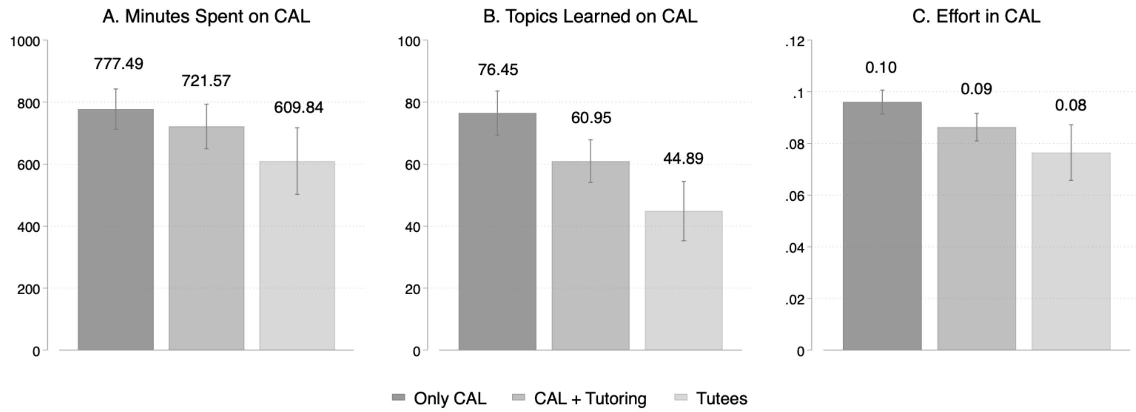
Notes: The blue line shows average baseline score for selected tutees.

Figure 1.8: Share of Correct Answers by Performance Groups



Notes: The 20th percentile or below threshold of performance is used as a counterfactual for tutees. "CAL+Tutoring Non-Tutees" shows trends between baseline and endline for students who did not receive the group tutoring intervention but were in classrooms where 20 percent of students did. "Tutees" are students who were assigned to the tutoring intervention in selected classrooms. "CAL-Only > 20th perc." are students in CAL-only classrooms who are above the 20th percentile of performance. "CAL Only ≤ 20th percentile" are students in CAL-only classrooms who were below the 20th percentile of performance. "Pure Control ≤ 20th percentile" are students below the 20th percentile performance threshold in control classrooms.

Figure 1.9: Platform Use by Groups



Notes: The CAL+Tutoring group excludes tutees.

Table 1.1: Balance Table

	Control		Treatment			
Variable	N	Mean/(SE)	N	Mean/(SE)	N	Difference
Female	669	0.552	664	0.566	1333	-0.015
		(0.019)		(0.019)		
Age	615	14.525	629	14.458	1244	0.067
		(0.033)		(0.033)		
Math test scores	620	6.281	631	6.442	1251	-0.162
		(0.108)		(0.110)		
Mother went to college	622	0.293	636	0.319	1258	-0.027
		(0.018)		(0.018)		
Father went to college	622	0.186	636	0.212	1258	-0.026
		(0.016)		(0.016)		
Mother went to high school	622	0.693	636	0.708	1258	-0.015
		(0.019)		(0.018)		
Father went to high school	622	0.564	636	0.558	1258	0.006
		(0.020)		(0.020)		
F-test of joint significance						0.749
F-test, number of observations						1237

Notes: *** p<0.01, ** p<0.05, * p<0.1. Math test scores are based on a short assessment administered at baseline.

Table 1.2: Attrition between Baseline, Midline, and Endline

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Midline Survey				Endline Survey			
	All	All	Treat	Control	All	All	Treat	Control
Treatment	0.005	0.017			0.029*	0.032*		
	(0.015)	(0.015)			(0.015)	(0.016)		
Student's age		-0.011	-0.024*	0.004		-0.029**	-0.037**	-0.022
		(0.011)	(0.014)	(0.018)		(0.012)	(0.015)	(0.020)
Student is female		-0.018	-0.006	-0.025		0.008	0.039*	-0.018
		(0.016)	(0.021)	(0.024)		(0.017)	(0.022)	(0.027)
Father is unemployed		-0.027	-0.036	-0.006		-0.009	0.030	-0.035
		(0.027)	(0.037)	(0.041)		(0.030)	(0.039)	(0.047)
Mother is unemployed		-0.024	0.005	-0.050**		-0.002	0.011	-0.007
		(0.017)	(0.023)	(0.025)		(0.019)	(0.024)	(0.029)
Baseline test score		0.001	0.000	0.004		0.006*	0.009**	0.004
		(0.003)	(0.004)	(0.005)		(0.003)	(0.004)	(0.005)
Repeated a school year		0.002	0.026	-0.023		0.004	0.035	-0.023
		(0.025)	(0.032)	(0.039)		(0.027)	(0.034)	(0.044)
Mother went to college		0.025	0.036	0.019		-0.015	-0.024	-0.004
		(0.018)	(0.024)	(0.028)		(0.020)	(0.025)	(0.032)
Mother went to high school		0.032*	0.011	0.045		0.024	0.012	0.034
		(0.019)	(0.026)	(0.029)		(0.021)	(0.027)	(0.033)
Father went to high school		-0.037**	0.007	-0.076***		-0.032*	-0.044*	-0.018
		(0.018)	(0.024)	(0.026)		(0.019)	(0.025)	(0.030)
Father went to college		-0.012	-0.016	-0.017		-0.007	0.008	-0.022
		(0.020)	(0.026)	(0.033)		(0.022)	(0.027)	(0.037)
Locus of control		0.068	0.103	0.028		0.008	-0.002	0.011
		(0.063)	(0.085)	(0.095)		(0.070)	(0.091)	(0.109)
Self-efficacy		-0.061	-0.099	-0.028		-0.068	-0.096	-0.057
		(0.049)	(0.066)	(0.073)		(0.054)	(0.071)	(0.083)
Spanish grade in 2022/23		0.002**	0.003*	0.002		-0.001	-0.001	-0.001
		(0.001)	(0.001)	(0.002)		(0.001)	(0.002)	(0.002)
Math grade in 2022/23		-0.002	-0.003	-0.002		0.003**	0.002	0.003
		(0.001)	(0.002)	(0.002)		(0.001)	(0.002)	(0.002)
Constant	0.916***	1.096***	1.286***	0.927***	0.900***	1.242***	1.327***	1.204***
	(0.011)	(0.188)	(0.251)	(0.290)	(0.011)	(0.208)	(0.266)	(0.331)
R-squared	0.000	0.037	0.043	0.055	0.003	0.033	0.051	0.033
N	1333	948	473	475	1333	948	473	475

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table 1.3: Treatment Effects on Math Assessments

	Midline Math Test		Endline Math Test	
	(1)	(2)	(3)	(4)
CAL Treatment (β_1)	0.206	0.167	0.313*	0.285**
	[0.179]	[0.180]	[0.057]	[0.033]
CAL x Tutoring Treatment (β_2)	-0.103	-0.066	-0.239	-0.194
	[0.619]	[0.656]	[0.214]	[0.208]
CAL + Tutoring Treatment (β_1 + β_2)	0.103	0.101	0.073	0.091
	[0.485]	[0.401]	[0.653]	[0.518]
Control Mean	0.00	0.00	0.00	0.00
Adjusted R^2	0.01	0.28	0.02	0.30
Observations	1225	1225	1218	1218
Covariates	No	Yes	No	Yes

Notes: *** p<0.01, ** p<0.05, * p<0.1. Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences towards math, self-reported grades for the previous year and parents' employment status. When available, we also control for the baseline level of the outcome variable. The p-values shown in brackets represent the probability of obtaining an estimated treatment effect at least as large in absolute value as the one observed in our experiment, under the sharp null hypothesis that the treatment effect is zero for all observations. Randomization inference is based on 2,000 random draws.

Table 1.4: Treatment Effects on Math Intrinsic Motivation

	Math Intrinsic Motivation Index	I enjoy learning math	I wish I did not have to study math	I like to solve math problems
	(1)	(2)	(3)	(4)
CAL Treatment (β_1)	-0.026	0.036	-0.002	-0.133**
	[0.594]	[0.732]	[0.841]	[0.022]
CAL x Tutoring Treatment (β_2)	0.102	0.071	-0.009	0.093
	[0.349]	[0.480]	[0.891]	[0.310]
CAL + Tutoring Treatment (β_1 + β_2)	0.076	0.106	-0.011	-0.041
	[0.578]	[0.223]	[0.987]	[0.395]
Adjusted R^2	-0.00	-0.00	-0.00	0.00
Observations	1194	1191	1179	1181

Notes: *** p<0.01, ** p<0.05, * p<0.1. Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences towards math, self-reported grades for the previous year and parents' employment status. When available, we also control for the baseline level of the outcome variable. The *p*-values shown in brackets represent the probability of obtaining an estimated treatment effect at least as large in absolute value as the one observed in our experiment, under the sharp null hypothesis that the treatment effect is zero for all observations. Randomization inference is based on 2,000 random draws.

Table 1.5: Treatment Effects on Math Preferences

	Student Likes Math			
	Midline Preferences		Endline Preferences	
	(1)	(2)	(3)	(4)
CAL Treatment (β_1)	-0.051	0.005	-0.021	0.020
	[0.560]	[0.944]	[0.832]	[0.809]
CAL x Tutoring Treatment (β_2)	-0.030	-0.038	0.023	0.024
	[0.819]	[0.701]	[0.852]	[0.835]
CAL + Tutoring Treatment (β_1 + β_2)	-0.081	-0.032	0.002	0.044
	[0.379]	[0.669]	[0.984]	[0.618]
Control Mean	0.00	0.00	0.00	0.00
Adjusted R^2	-0.00	0.32	-0.00	0.24
Observations	1218	1218	1210	1210
Covariates	No	Yes	No	Yes

Notes: *** p<0.01, ** p<0.05, * p<0.1. Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences towards math, self-reported grades for the previous year and parents' employment status. When available, we also control for the baseline level of the outcome variable. The p-values shown in brackets represent the probability of obtaining an estimated treatment effect at least as large in absolute value as the one observed in our experiment, under the sharp null hypothesis that the treatment effect is zero for all observations. Randomization inference is based on 2,000 random draws.

Table 1.6: Treatment Effects on Absence Rates

	Share of Days Absent	
	(1)	(2)
CAL Treatment (β_1)	-0.007	-0.002
	[0.697]	[0.697]
CAL x Tutoring Treatment (β_2)	0.010	0.010
	[0.535]	[0.535]
CAL + Tutoring Treatment ($\beta_1 + \beta_2$)	0.008	0.008
	[0.658]	[0.658]
Control Mean	0.09	0.09
Adjusted R^2	0.00	0.00
Observations	1254	1254
Covariates	No	Yes

Notes: *** p<0.01, ** p<0.05, * p<0.1. Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences towards math, self-reported grades for the previous year and parents' employment status. When available, we also control for the baseline level of the outcome variable. The *p*-values shown in brackets represent the probability of obtaining an estimated treatment effect at least as large in absolute value as the one observed in our experiment, under the sharp null hypothesis that the treatment effect is zero for all observations. Randomization inference is based on 2,000 random draws. Attendance measures are specific to math classes.

Table 1.7: Treatment Effects on School Dropout

	School Dropout	
	(1)	(2)
CAL Treatment (β_1)	-0.019	-0.010
	[0.442]	[0.603]
CAL x Tutoring Treatment (β_2)	-0.009	-0.016
	[0.806]	[0.564]
CAL + Tutoring Treatment ($\beta_1 + \beta_2$)	-0.028	-0.026
	[0.263]	[0.170]
Control Mean	0.08	0.08
Adjusted R^2	0.00	0.06
Observations	1333	1333
Covariates	No	Yes

Notes: *** p<0.01, ** p<0.05, * p<0.1. Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences towards math, self-reported grades for the previous year and parents' employment status. When available, we also control for the baseline level of the outcome variable. The *p*-values shown in brackets represent the probability of obtaining an estimated treatment effect at least as large in absolute value as the one observed in our experiment, under the sharp null hypothesis that the treatment effect is zero for all observations. Randomization inference is based on 2,000 random draws.

Table 1.8: Treatment Effects on School Grades

	Math Grades		Spanish Grades	
	(1)	(2)	(3)	(4)
Panel A: Unconditional School Grades				
CAL Treatment (β_1)	-0.141	-0.153	0.012	-0.015
	[0.311]	[0.262]	[0.909]	[0.890]
CAL x Tutoring Treatment (β_2)	-0.095	-0.050	-0.072	-0.012
	[0.597]	[0.743]	[0.722]	[0.929]
CAL + Tutoring Treatment (β_1 + β_2)	-0.235	-0.203	-0.060	-0.027
	[0.102]	[0.146]	[0.599]	[0.779]
Adjusted R^2	0.01	0.31	-0.00	0.32
Observations	1247	1247	1247	1247
Covariates	No	Yes	No	Yes
Panel B: Conditional School Grades				
CAL Treatment (β_1)	-0.188	-0.202	0.029	0.006
	[0.227]	[0.158]	[0.812]	[0.959]
CAL x Tutoring Treatment (β_2)	-0.126	-0.069	-0.063	-0.009
	[0.520]	[0.666]	[0.772]	[0.953]
CAL + Tutoring Treatment (β_1 + β_2)	-0.313	-0.271	-0.034	-0.003
	[0.050]	[0.066]	[0.792]	[0.980]
Adjusted R^2	0.02	0.37	-0.00	0.31
Observations	1247	1247	1247	1247
Covariates	No	Yes	No	Yes

Notes: *** p<0.01, ** p<0.05, * p<0.1. Panel A reports results for unconditional school grades, calculated as the average of all quarterly grades available for the 2023-2024 school year. In contrast, Panel B shows conditional school grades, which reflect the final grade reported at the end of the school year. Unlike unconditional grades, conditional grades are only available for students who remained enrolled through the end of the academic year. Covariates include students' gender, age, baseline math test scores, and proxies for confidence and preferences towards math, along with self-reported grades for the previous year and parents' employment status. When available, we also control for the baseline level of the outcome variable. The p-values shown in brackets represent the probability of obtaining an estimated treatment effect at least as large in absolute value as the one observed in our experiment, under the sharp null hypothesis that the treatment effect is zero for all observations. Randomization inference is based on 2,000 random draws.

Table 1.9: Treatment Effects on Math Self-Beliefs and Self-Confidence

	Midline Math Self-Beliefs		Endline Math Self-Beliefs	
	(1)	(2)	(3)	(4)
Panel A: Math Self-Beliefs (SD)				
CAL Treatment (β_1)	0.117	0.127	0.119	0.134
	[0.237]	[0.186]	[0.156]	[0.100]
CAL x Tutoring Treatment (β_2)	-0.009	0.020	0.007	0.035
	[0.941]	[0.872]	[0.938]	[0.725]
CAL + Tutoring Treatment ($\beta_1 + \beta_2$)	0.108	0.147	0.126	0.169**
	[0.254]	[0.111]	[0.133]	[0.038]
Adjusted R^2	0.00	0.15	0.00	0.14
Observations	1205	1205	1188	1188
Covariates	No	Yes	No	Yes
Panel B: Math Self-Confidence Index				
CAL Treatment (β_1)	-0.016	-0.007	-0.010	-0.002
	[0.274]	[0.605]	[0.603]	[0.923]
CAL x Tutoring Treatment (β_2)	0.011	0.014	-0.005	-0.000
	[0.599]	[0.368]	[0.820]	[0.999]
CAL + Tutoring Treatment ($\beta_1 + \beta_2$)	-0.005	0.008	-0.014	-0.002
	[0.713]	[0.533]	[0.422]	[0.918]
Control Mean	0.66	0.66	0.64	0.64
Adjusted R^2	0.00	0.34	-0.00	0.28
Observations	1225	1225	1206	1206
Covariates	No	Yes	No	Yes

Notes: *** p<0.01, ** p<0.05, * p<0.1. Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences towards math, self-reported grades for the previous year and parents' employment status. When available, we also control for the baseline level of the outcome variable. The p-values shown in brackets represent the probability of obtaining an estimated treatment effect at least as large in absolute value as the one observed in our experiment, under the sharp null hypothesis that the treatment effect is zero for all observations. Randomization inference is based on 2,000 random draws. The outcome variable in Panel A is a standardized measure of students' self-reported expected performance on a math test. The outcome variable in Panel B is a standardized index based on students' answers to the following items: 1) When teachers teach me a new math topic, I learn it without much difficulty; 2) When I take a math test, I am confident that I can answer the questions well; 3) I can help my classmates with math homework that they don't understand; 4) I feel more capable as I learn more things in math. Answers ranged 1 (Never) to 4 (Always).

Table 1.10: Treatment Effects on Math Self-Concept

	Math Self-Concept Index	I usually do well in math	Math is harder for me than for many of my classmates
	(1)	(2)	(3)
CAL Treatment (β_1)	-0.008	-0.005	0.032
	[0.844]	[0.947]	[0.486]
CAL x Tutoring Treatment (β_2)	0.011	-0.003	-0.093
	[0.927]	[0.846]	[0.292]
CAL + Tutoring Treatment (β_1 + β_2)	0.003	-0.008	-0.060
	[0.832]	[0.648]	[0.527]
Adjusted R^2	-0.00	-0.00	-0.00
Observations	1184	1182	1175

Notes: *** p<0.01, ** p<0.05, * p<0.1. Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences towards math, self-reported grades for the previous year and parents' employment status. When available, we also control for the baseline level of the outcome variable. The *p*-values shown in brackets represent the probability of obtaining an estimated treatment effect at least as large in absolute value as the one observed in our experiment, under the sharp null hypothesis that the treatment effect is zero for all observations. Randomization inference is based on 2,000 random draws.

Table 1.11: Treatment Effects on Math Overconfidence

	Midline Overconfidence		Endline Overconfidence	
	(1)	(2)	(3)	(4)
CAL Treatment (β_1)	0.018	0.032	-0.002	0.008
	[0.698]	[0.499]	[0.956]	[0.844]
CAL x Tutoring Treatment (β_2)	0.026	0.029	0.042	0.050
	[0.645]	[0.609]	[0.331]	[0.269]
CAL + Tutoring Treatment (β_1 + β_2)	0.044	0.061	0.040	0.058
	[0.351]	[0.203]	[0.289]	[0.132]
Adjusted R^2	-0.00	0.05	-0.00	0.05
Observations	1205	1205	1188	1188
Covariates	No	Yes	No	Yes

Notes: *** p<0.01, ** p<0.05, * p<0.1. Notes: *** p<0.01, ** p<0.05, * p<0.1. The outcome variable takes the value of 1 if the teacher switched classes after program assignment, and 0 otherwise. Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences towards math, self-reported grades for the previous year and parents' employment status. When available, we also control for the baseline level of the outcome variable. The *p*-values shown in brackets represent the probability of obtaining an estimated treatment effect at least as large in absolute value as the one observed in our experiment, under the sharp null hypothesis that the treatment effect is zero for all observations. Randomization inference is based on 2,000 random draws.

Table 1.12: Treatment Effects on Self-Efficacy Index

	Midline Self-Efficacy Index		Endline Self-Efficacy Index	
	(1)	(2)	(3)	(4)
CAL Treatment (β_1)	-0.016	-0.018	-0.009	-0.012
	[0.284]	[0.151]	[0.534]	[0.361]
CAL x Tutoring Treatment (β_2)	0.017	0.020	-0.000	0.006
	[0.368]	[0.126]	[0.986]	[0.737]
CAL + Tutoring Treatment (β_1 + β_2)	0.001	0.002	-0.010	-0.006
	[0.941]	[0.875]	[0.507]	[0.625]
Adjusted R^2	0.00	0.25	-0.00	0.13
Observations	1217	1217	1194	1194
Covariates	No	Yes	No	Yes

Notes: *** p<0.01, ** p<0.05, * p<0.1. Notes: *** p<0.01, ** p<0.05, * p<0.1. Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences towards math, self-reported grades for the previous year and parents' employment status. When available, we also control for the baseline level of the outcome variable. The p-values shown in brackets represent the probability of obtaining an estimated treatment effect at least as large in absolute value as the one observed in our experiment, under the sharp null hypothesis that the treatment effect is zero for all observations. Randomization inference is based on 2,000 random draws.

Table 1.13: Treatment Effects on Grit Index

	Grit Index	
	(1)	(2)
CAL Treatment (β_1)	0.011	0.006
	[0.430]	[0.676]
CAL x Tutoring Treatment (β_2)	0.001	0.001
	[0.965]	[0.968]
CAL + Tutoring Treatment (β_1 + β_2)	0.012	0.006
	[0.361]	[0.653]
Adjusted R^2	-0.00	0.03
Observations	1333	1333
Covariates	No	Yes

Notes: *** p<0.01, ** p<0.05, * p<0.1. Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences towards math, self-reported grades for the previous year and parents' employment status. When available, we also control for the baseline level of the outcome variable. The *p*-values shown in brackets represent the probability of obtaining an estimated treatment effect at least as large in absolute value as the one observed in our experiment, under the sharp null hypothesis that the treatment effect is zero for all observations. Randomization inference is based on 2,000 random draws. The outcome is a standardized variable based on students' answers to the following items: 1) The schoolwork I like the most is the one that makes me think, 2) If I think I will lose in a game, I do not want to continue playing, 3) If I set a goal and see that it's harder than I though I lose interest, and 4) I work hard on tasks. Answer options ranged from 1 (Doesn't seem like me at all) to 5 (Very much like me).

Table 1.14: Treatment Effects on Logic Task

	Logic Task: Easy	Logic Task: difficult	Logic Task: Give Up
	(1)	(2)	(3)
CAL Treatment (β_1)	0.087	-0.094	0.007
	[0.075]	[0.019]	[0.730]
CAL x Tutoring Treatment (β_2)	-0.037	0.023	0.014
	[0.651]	[0.851]	[0.720]
CAL + Tutoring Treatment ($\beta_1 + \beta_2$)	0.050	-0.071	0.021
	[0.201]	[0.043]	[0.417]
Control Mean	0.43	0.38	0.19
Adjusted R^2	0.00	0.07	-0.00
Observations	1211	1211	1211

Notes: *** p<0.01, ** p<0.05, * p<0.1. Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences towards math, self-reported grades for the previous year and parents' employment status. When available, we also control for the baseline level of the outcome variable. The p-values shown in brackets represent the probability of obtaining an estimated treatment effect at least as large in absolute value as the one observed in our experiment, under the sharp null hypothesis that the treatment effect is zero for all observations. Randomization inference is based on 2,000 random draws.

Table 1.15: Treatment Effects on Locus of Control

	Locus of Control	
	(1)	(2)
CAL Treatment (β_1)	0.010	0.004
	[0.258]	[0.585]
CAL x Tutoring Treatment (β_2)	-0.025*	-0.020*
	[0.056]	[0.056]
CAL + Tutoring Treatment ($\beta_1 + \beta_2$)	-0.014	-0.015*
	[0.120]	[0.056]
Control Mean	0.67	0.67
Adjusted R^2	0.00	0.11
Observations	1196	1196
Covariates	No	Yes

Notes: *** p<0.01, ** p<0.05, * p<0.1. Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences towards math, self-reported grades for the previous year and parents' employment status. When available, we also control for the baseline level of the outcome variable. The p-values shown in brackets represent the probability of obtaining an estimated treatment effect at least as large in absolute value as the one observed in our experiment, under the sharp null hypothesis that the treatment effect is zero for all observations. Randomization inference is based on 2,000 random draws.

Table 1.16: Treatment Effects on Student Time Use

	Hours spent on:			
	Academic Activities	Work	Entertainment	Sleep
	(1)	(2)	(3)	(4)
CAL Treatment (β_1)	0.150	-0.163	0.159	0.001
	[0.607]	[0.773]	[0.499]	[0.956]
CAL x Tutoring Treatment (β_2)	-0.191	-0.375	-0.223	0.105
	[0.503]	[0.697]	[0.395]	[0.556]
CAL + Tutoring Treatment ($\beta_1 + \beta_2$)	-0.041	-0.538	-0.223	0.106
	[0.838]	[0.277]	[0.539]	[0.385]
Control Mean	7.90	8.61	7.55	5.66
Adjusted R^2	-0.00	-0.00	-0.00	-0.00
Observations	1194	1182	1185	1115

Notes: *** p<0.01, ** p<0.05, * p<0.1. Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences towards math, self-reported grades for the previous year and parents' employment status. When available, we also control for the baseline level of the outcome variable. The p-values shown in brackets represent the probability of obtaining an estimated treatment effect at least as large in absolute value as the one observed in our experiment, under the sharp null hypothesis that the treatment effect is zero for all observations. Randomization inference is based on 2,000 random draws.

Table 1.17: Treatment Effects on Student Depression

	Depression Index	
	(1)	(2)
CAL Treatment (β_1)	0.012	0.009
	[0.286]	[0.306]
CAL x Tutoring Treatment (β_2)	-0.003	-0.003
	[0.847]	[0.755]
CAL + Tutoring Treatment (β_1 + β_2)	0.009	0.006
	[0.396]	[0.511]
Adjusted R^2	-0.00	0.08
Observations	1205	1205
Covariates	No	Yes

Notes: *** p<0.01, ** p<0.05, * p<0.1. Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences towards math, self-reported grades for the previous year and parents' employment status. When available, we also control for the baseline level of the outcome variable. The *p*-values shown in brackets represent the probability of obtaining an estimated treatment effect at least as large in absolute value as the one observed in our experiment, under the sharp null hypothesis that the treatment effect is zero for all observations. Randomization inference is based on 2,000 random draws. The outcome variable is a standardized index based on students' answers to the following items: 1) I am happy, 2) I worry a lot, 3) I feel sad, 4) I get upset quickly, 5) I often think I did something wrong, 6) It's hard for me to concentrate, 7) I feel lonely, and 8) I am not happy. Answers range from 1 (Strongly disagree) to 4 (Strongly agree).

Table 1.18: Treatment Effects on Math Anxiety

	Math Anxiety Index	I'm worried of having bad grades in math	I get nervous before math tests
	(1)	(2)	(3)
CAL Treatment (β_1)	0.135	0.033	0.060
	[0.078]	[0.265]	[0.049]
CAL x Tutoring Treatment (β_2)	-0.123	-0.016	-0.061
	[0.229]	[0.657]	[0.115]
CAL + Tutoring Treatment (β_1 + β_2)	0.012	0.017	-0.001
	[0.229]	[0.598]	[0.819]
Control Mean	0.00	0.77	0.71
Adjusted R^2	0.00	0.00	0.00
Observations	1182	1199	1183

Notes: *** p<0.01, ** p<0.05, * p<0.1. Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences towards math, self-reported grades for the previous year and parents' employment status. When available, we also control for the baseline level of the outcome variable. The p-values shown in brackets represent the probability of obtaining an estimated treatment effect at least as large in absolute value as the one observed in our experiment, under the sharp null hypothesis that the treatment effect is zero for all observations. Randomization inference is based on 2,000 random draws.

Table 1.19: Treatment Effects on Teacher Behavior

	Teacher Behavior	
	(1)	(2)
Panel A: Time Mismanagement Index		
CAL Treatment (β_1)	-0.082	-0.030
	[0.556]	[0.801]
CAL x Tutoring Treatment (β_2)	-0.145	-0.146
	[0.382]	[0.301]
CAL + Tutoring Treatment ($\beta_1 + \beta_2$)	-0.145	-0.146
	[0.382]	[0.301]
Adjusted R^2	0.01	0.12
Observations	1148	1148
Covariates	No	Yes
Panel B: Teacher Engagement Index		
CAL Treatment (β_1)	-0.181	-0.144
	[0.102]	[0.112]
CAL x Tutoring Treatment (β_2)	0.044	0.041
	[0.773]	[0.763]
CAL + Tutoring Treatment ($\beta_1 + \beta_2$)	-0.137	-0.103
	[0.220]	[0.249]
Adjusted R^2	0.00	0.12
Observations	1151	1151
Covariates	No	Yes

Notes: *** p<0.01, ** p<0.05, * p<0.1. Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences towards math, self-reported grades for the previous year and parents' employment status. The p-values shown in brackets represent the probability of obtaining an estimated treatment effect at least as large in absolute value as the one observed in our experiment, under the sharp null hypothesis that the treatment effect is zero for all observations. Randomization inference is based on 2,000 random draws.

Table 1.20: Treatment Effects on Teaching Effectiveness Index

	Teacher Effectiveness Index		Teacher Ineffectiveness Index	
	(1)	(2)	(3)	(4)
CAL Treatment (β_1)	0.055	0.059	0.068	0.083
	[0.606]	[0.579]	[0.480]	[0.390]
CAL x Tutoring Treatment (β_2)	-0.144	-0.122	-0.155	-0.148
	[0.294]	[0.359]	[0.245]	[0.245]
CAL + Tutoring Treatment ($\beta_1 + \beta_2$)	-0.089	-0.063	-0.087	-0.066
	[0.403]	[0.555]	[0.379]	[0.530]
Adjusted R^2	0.00	0.04	0.00	0.03
Observations	1161	1161	1165	1165
Covariates	No	Yes	No	Yes

Notes: *** p<0.01, ** p<0.05, * p<0.1. Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences towards math, self-reported grades for the previous year and parents' employment status. The p-values shown in brackets represent the probability of obtaining an estimated treatment effect at least as large in absolute value as the one observed in our experiment, under the sharp null hypothesis that the treatment effect is zero for all observations. Randomization inference is based on 2,000 random draws

Table 1.21: Treatment Effects on Test Scores by Content Level

	Below Grade-Level		At Grade-Level	
	(1)	(2)	(3)	(4)
CAL Treatment (β_1)	0.282**	0.262**	0.243	0.214*
	[0.034]	[0.028]	[0.102]	[0.073]
CAL x Tutoring Treatment (β_2)	-0.163	-0.127	-0.262	-0.220
	[0.304]	[0.379]	[0.153]	[0.118]
CAL + Tutoring Treatment (β_1 + β_2)	0.119	0.135	-0.019	-0.006
	[0.400]	[0.340]	[0.901]	[0.955]
Adjusted R^2	0.01	0.22	0.01	0.24
Observations	1218	1218	1217	1217
Covariates	No	Yes	No	Yes

Notes: *** p<0.01, ** p<0.05, * p<0.1. Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences towards math, self-reported grades for the previous year and parents' employment status. When available, we also control for the baseline level of the outcome variable. The *p*-values shown in brackets represent the probability of obtaining an estimated treatment effect at least as large in absolute value as the one observed in our experiment, under the sharp null hypothesis that the treatment effect is zero for all observations. Randomization inference is based on 2,000 random draws.

Table 1.22: Treatment Effects on Test Scores by Math Area

	Midline			Endline			
	Algebra	Numbers	Data and Prob	Algebra	Numbers	Data and Prob	Geometry
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
CAL Treatment (β_1)	0.195*	0.093	-0.011	0.214*	0.317**	0.028	0.133
	[0.100]	[0.376]	[0.929]	[0.088]	[0.022]	[0.679]	[0.229]
CAL x Tutoring (β_2)	-0.091	0.011	-0.062	-0.220	-0.106	0.056	-0.191
	[0.503]	[0.910]	[0.460]	[0.147]	[0.451]	[0.693]	[0.101]
CAL x Tutoring ($\beta_1 + \beta_2$)	0.104	0.104	-0.073	-0.006	0.210	0.084	-0.058
	[0.380]	[0.376]	[0.380]	[0.147]	[0.149]	[0.382]	[0.490]
Adjusted R^2	0.01	0.19	-0.00	0.01	0.02	-0.00	0.00
Observations	1225	1225	1220	1217	1217	1218	1218

Notes: *** p<0.01, ** p<0.05, * p<0.1. Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences towards math, self-reported grades for the previous year and parents' employment status. When available, we also control for the baseline level of the outcome variable. The *p*-values shown in brackets represent the probability of obtaining an estimated treatment effect at least as large in absolute value as the one observed in our experiment, under the sharp null hypothesis that the treatment effect is zero for all observations. Randomization inference is based on 2,000 random draws.

Table 1.23: Treatment Effects on Test Scores by Knowledge Domain

	Midline			Endline		
	Knowing	Applying	Reasoning	Knowing	Applying	Reasoning
	(1)	(2)	(3)	(4)	(5)	(6)
CAL Treatment (β_1)	0.198	0.075	0.075	0.311***	0.164	0.211**
	[0.170]	[0.346]	[0.380]	[0.007]	[0.182]	[0.042]
CAL x Tutoring (β_2)	0.009	-0.064	-0.156	-0.213	-0.097	-0.089
	[0.916]	[0.547]	[0.160]	[0.120]	[0.456]	[0.372]
CAL + Tutoring (β_1 + β_2)	0.009	-0.064	-0.156	-0.213	-0.097	-0.089
	[0.916]	[0.547]	[0.160]	[0.120]	[0.456]	[0.372]
Adjusted R^2	0.01	0.18	0.00	0.02	0.00	0.01
Observations	1225	1225	1224	1218	1217	1218

Notes: *** p<0.01, ** p<0.05, * p<0.1. Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences towards math, self-reported grades for the previous year and parents' employment status. When available, we also control for the baseline level of the outcome variable. The *p*-values shown in brackets represent the probability of obtaining an estimated treatment effect at least as large in absolute value as the one observed in our experiment, under the sharp null hypothesis that the treatment effect is zero for all observations. Randomization inference is based on 2,000 random draws.

Table 1.24: Treatment Effects on Math Assessments by Levels of Locus of Control

	Midline Math Test		Endline Math Test	
	(1)	(2)	(3)	(4)
CAL Treatment, Low LOC	0.234	0.179	0.295	0.269
	[0.501]	[0.503]	[0.485]	[0.495]
CAL Treatment, High LOC	0.178	0.154	0.332**	0.303**
	[0.220]	[1.000]	[0.033]	[0.022]
CAL+ Tutoring Treatment, High LOC	0.273*	0.260**	0.200	0.204
	[0.068]	[0.033]	[0.196]	[0.122]
CAL + Tutoring Treatment, Low LOC	-0.006	-0.014	0.002	0.017
	[0.973]	[0.903]	[0.992]	[0.897]
Control Mean	0.00	0.00	0.00	0.00
Adjusted R^2	0.03	0.28	0.05	0.31
Observations	1225	1225	1218	1218
Covariates	No	Yes	No	Yes

Notes: *** p<0.01, ** p<0.05, * p<0.1. Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences towards math, self-reported grades for the previous year and parents' employment status. When available, we also control for the baseline level of the outcome variable. The p-values shown in brackets represent the probability of obtaining an estimated treatment effect at least as large in absolute value as the one observed in our experiment, under the sharp null hypothesis that the treatment effect is zero for all observations. Randomization inference is based on 2,000 random draws.

Chapter 2. Different Paths to College Success: The Impact of Massachusetts' Charter Schools on College Trajectories[1]

2.1 Introduction

Despite the recent decline in the college wage premium, college graduates still outearn their peers with only a high-school diploma by 75 percent (Bengali et al., 2023; Autor et al., 2023). With much policy and research in the United States focused on college access, we know less about how K–12 educational experiences contribute to college success (Dynarski et al., 2023). This paper uses application lotteries to show the causal effects of one K–12 educational intervention—charter schools—on college preparation, enrollment, and graduation.

Charter schools are autonomously operated public schools with oversight, curricular, budgetary, and hiring independence from traditional school districts. They are authorized by a state-empowered entity, undergo periodic review, and may be subject to closure. When oversubscribed, charter schools admit students via randomized admissions lotteries. Charter schools are not monolithic in character. Many urban charter schools feature longer school days and school years, a culture of high expectations, frequent teacher observations and feedback, data-driven instruction, use of tutoring, and strict discipline—practices which are often referred to as "No Excuses" (Angrist et al., 2013; Dobbie and Fryer, 2013). In recent years, many of these schools have moved away from this label and some of the associated practices (Torres, 2022). Other charter schools operate on the basis of a greater range of educational models and include project-based learning schools, themed schools (e.g. arts, language, STEM, culture), Montessori schools, and personalized learning schools.

[1] Co-authored with Sarah Cohodes.

This paper builds on Angrist et al. (2013) (APW), which examines the effects of Massachusetts charter schools on test scores across urban and nonurban areas. APW find that urban charter schools generate large test score gains whereas nonurban charters have null or negative effects. In our shared Massachusetts sample, at the time students were enrolled in the 2000's and 2010's, the urban charter schools mostly adhered to "No Excuses" practices and served a primarily minority and economically disadvantaged population. The nonurban schools did not, embracing alternative charter school models—in particular, project-based learning—and serving primarily white children. APW find that the different practices and student bodies help account for the different test scores trajectories, aligning with the existing literature on charter schools.

With a longer time horizon and more cohorts and schools, we return to APW's diverse sample of charter schools and report several novel findings. First, we replicate their test score results, finding that urban charters boost standardized test scores and nonurban charters do not. Next, we find that urban and nonurban charters both accelerate college preparation but via different means. Urban charters increase Advanced Placement (AP) and SAT test-taking and scores and completion of a college-ready curriculum, but increase time to high-school graduation. Nonurban charters decrease AP test-taking but boost completion of a college-ready curriculum.

Turning to college, both urban and nonurban charter schools boost four-year college enrollment, by 7.9 and 9.8 percentage points, respectively. Regarding college graduation, we find that attending an urban charter school raises attainment of any degree by 4.6 percentage points from the comparison mean of 24 percent and BA completion by 4.2 percentage points from a comparison of 22 percent. Nonurban charter schools increase attainment of any degree by 11.2 percentage points from a comparison mean of 52 percent and BA attainment by 11.9 percentage points over a comparison rate of 47 percent. In short, both urban and nonurban charter schools lift degree attainment, but nonurban charter schools—the same schools that do not boost test scores—induce very large gains, more than twice as large as the college graduation boost from urban charters.

Our test score findings are consistent with the many lottery-based studies that have shown attending urban charter schools increases students' test scores (see Cohodes and Roy (2024) for a summary of this research) and the more limited lottery-based evidence on nonurban charter schools, which shows mixed impacts on test scores, with findings of small positive effects (Dynarski et al., 2018) and, in other cases, null or negative effects (Gleason et al., 2010; Angrist et al., 2013). Observational estimates of charter school impacts appear to confirm the lottery-based evidence that urban charter schools boost test scores while nonurban charters do not (see Cohodes and Parham (2021) for an overview).

The test score evidence has led some to conclude that charter schools are most successful in urban contexts when they adopt "No Excuses" practices (Chabrier et al., 2016; Epple et al., 2016; Cohodes and Parham, 2021). This conclusion is bolstered by lottery-based evidence showing that urban charters that boost test scores also boost college preparation and enrollment and even shape non–test score outcomes such as voting and risky behavior (Angrist et al., 2016; Dobbie and Fryer, 2015; Wong et al., 2014; Davis and Heller, 2019; Cohodes and Feigenbaum, 2021; Reber et al., 2023; Demers et al., 2017). However, there is much less evidence on nonurban charter schools and nontest outcomes. Lottery-based evidence on college graduation comes from a broad sample of charter schools in a federally funded national evaluation of charter schools (Gleason et al., 2010) extended to college outcomes (Place and Gleason, 2019), which finds no impact on college enrollment or graduation and no relationship between test scores and college outcomes. This is perhaps because their sample only consists of middle schools, whereas most of the schools in our sample offer high-school grades. Evidence from the mostly-urban KIPP schools is consistent with the idea that high-school grades are important. Demers et al. (2017) find no college enrollment or graduation boost from a lottery-based evaluation of KIPP middle schools, but when they add an instrumental variables approach to account for attendance at a KIPP high school, they find large college gains. Dobbie and Fryer (2020) use propensity score matching to show that "No Excuses" charter schools improve test scores and four-year college enrollment whereas "other" charter schools decrease both.[2]

This paper makes two main contributions. First, we add to the evidence on charter schools by presenting lottery-based estimates of their impacts on college graduation from a diverse sample of schools. The findings expand our as yet limited knowledge of the impact of different charter school models on college outcomes. Second, we demonstrate that charter school test score effects do not always align with the schools' impacts on students' life trajectories. Standardized test scores provide a useful but limited measure of student learning. Jackson (2018) and Jackson et al. (2020) highlight that teacher and school effects on test scores and student behavior separably contribute to longer-term outcomes. Nevertheless, researchers often use standardized test scores as a proxy for other outcomes that we care about (Krueger, 2003; Chetty et al., 2011; Hanushek, 2011; Chetty et al., 2014c; Ganimian et al., 2021), but had we done so for our sample of Massachusetts charter schools, we would have come to the wrong conclusion about the schools' impacts on attainment. Now that sufficient time has passed for APW's sample of students to have completed their education, we can measure their longer-term outcomes directly and do so in the remainder of this paper.

2.2 Data and Descriptive Statistics

2.2.1 Data and Sample

Massachusetts charter school records from randomized admissions lotteries in 2002–2014, corresponding to cohorts projected to graduate high school in 2006–2018, form the basis of our investigation into charter school impacts (Massachusetts Charter School Lottery Records, 2002–2014). We include schools with admission in the middle-school grades or later, as students admitted for elementary school are too young to observe longer-term outcomes. Our sample, based on APW's but augmented by a few additional schools, includes 15 urban charter schools and 9 nonurban charter schools. We define schools as urban if they are in towns where the school district participated in

[2]Observational work from Florida shows that charter schools initially decrease scores (Sass, 2006) but increase college persistence (Sass et al., 2016), but it does not differentiate by location type or model.

the Massachusetts Urban Superintendents Network and as nonurban otherwise. The sample covers all Massachusetts charter schools that offered admission for middle- or high-school grades at the time of the initial lottery record collection (2009–2011) and for which there are records of lotteries with more applicants than seats available (Appendix Table B.1). All students in the sample are old enough to be observed 5 years after their projected high-school graduation, with one fewer cohort available at 6 years after projected high-school graduation.

The lottery records include students' names and dates of birth alongside lottery information (application grade, sibling status, town of residence, admissions offers, and waitlist status). We use the lotteries for entry grades, as these have the greatest number of open seats and a standard open admission process, and exclude guaranteed-admission siblings and non-randomized late and out-of-area applicants. We create indicators both for admission on the day of the lottery (initial offers) and offers extended from the randomized waitlist (waitlist offers). The sample includes students present in the Massachusetts data at baseline, excluding students who applied to charter schools from private schools.

We use name, date of birth, town of residence, and application cohort to match the lottery records to state administrative data from the Massachusetts Department of Elementary and Secondary Education (DESE) (DESE, 2002-2023). These records include student information such as school enrollment, gender, race, special education status, English learner status, subsidized lunch status, days of attendance, suspensions, and high-school graduation status from the Student Information Management System (SIMS), course enrollment from the Student Course System (SCS), and achievement scores from the Massachusetts Comprehensive Assessment System (MCAS). DESE also provided information on AP and SAT exams from the College Board and college records from the National Student Clearinghouse (NSC).

Key outcomes include MCAS scores two years after the lottery, APs and SATs, MassCore curriculum completion, high-school graduation, college enrollment, and degree attainment. We standardize MCAS scores by subject, grade, and year to have mean zero and standard deviation one for the entire state.[3] AP and SAT outcomes are available for the class of 2007 and later. SAT and AP test participation indicators are marked as zeroes for students who were present in 11th or 12th grade but did not take each exam and missing if not present. MassCore completion is an indicator that the student has completed college-ready high-school curriculum, as defined by the state for the 2008 cohort and forward.[4] High-school graduation and MassCore completion is measured for students that appear in 9th grade. Indicators for IB and calculus course-taking come from the SCS but are only available for the 2015 and later cohorts. We report college outcomes within timeframes of expected high-school graduation, where the expected high-school graduation year is based on the year and grade of the application lottery. Thus, an outcome such as bachelor's attainment within 6 years indicates that a student obtained a bachelor's within 6 years of her expected high-school graduation based on when she applied to a charter school. We mark college indicators as zeroes if student information was sent to the NSC but no college records were returned, otherwise they are marked as missing.

2.2.2 Schools

In addition to diverging from their traditional public school counterparts, urban and nonurban charter schools diverge in their characteristics and practices from each other. Table 2.1 compares school characteristics for the charter schools in the lottery sample and other public schools. We measure school characteristics in the early 2010's, the point at which most students in our sample matriculated. Urban charter schools have the lowest share of teachers with formal credentials (59 percent licensed in their subject), followed by nonurban charters (71 percent), in contrast to public schools, where almost all teachers are licensed in their subject. All urban charter schools

[3]MCAS scores exclude middle-school scores from 2015 and 2016, when districts could select the MCAS or PARCC exam.

[4]The MassCore curriculum entails completing 4 years of math coursework, 4 years of ELA, 3 years of science, 3 years of history, 2 years of a world language, 1 year of arts and 5 additional units of core courses. The indicator is reported by school districts to the state.

receive federal Title 1 funds for serving a high-poverty student body, as do about two-thirds of nonurban charters. Among traditional schools, 77 percent of urban schools and 41 percent of nonurban schools receive Title 1. The student–teacher ratio is lower in charter schools (12:1 or 11:1 in charters and 14:1 or 13:1 in other public schools). Charter schools are small schools, with approximately 430 students per school. This compares to 663 students at urban traditional schools and 2,271 students per school in nonurban areas, which often have large comprehensive high schools. Per-pupil expenditures (in 2014 dollars) are slightly higher in urban charters ($16,250) than in urban traditional schools ($15,661) and lower in nonurban charters ($11,982) than in nonurban traditional schools ($14,411). Urban areas have higher disciplinary rates. However, relative to traditional public schools, urban charters use discipline more, whereas nonurban charters have fewer disciplinary incidents.

For the charter schools only, we have responses to a survey on school practices (Panel B). Urban charters have longer school days and school years, use tutoring, frequent teacher observations, and frequent checks for student understanding, and have a culture of high expectations. Two of the urban schools are affiliated with multi-state charter management organizations associated with No Excuses practices (KIPP and Uncommon Schools). Nonurban charters are less likely to deploy these practices, though half of them use frequent checks for student understanding and 75 percent use differentiated instruction (even higher than the 69 percent for urban charters). They are more likely to use project-based learning (63 percent versus 23 percent for urban charters). One of the nonurban schools is associated with the Coalition of Essential Schools, which focuses on individualized learning and civic contributions; another nonurban charter is associated with Expeditionary Learning, which emphasizes real-world projects and active learning;[5] and another nonurban school is an International Baccalaureate (IB) school, following a rigorous college-prep curriculum focused on critical thinking. Finally, one nonurban school focuses on performing arts. In all, 4 of 9 nonurban campuses have an explicit non-No Excuses affiliation or theme.

[5] One urban school also follows this model.

2.2.3 Students

Table 2.1, Panel C, also presents descriptive statistics for lottery applicants (Columns 1 and 2) and students who attended public schools in Massachusetts in 9th grade and were projected to graduate between 2006 and 2018 (Columns 3 and 4). We see important differences across urban and nonurban areas. In urban areas, Black and Latino/a students comprise 20 and 32 percent of the public school student population, respectively, and 53 and 28 percent of lottery applicants. Sixty-four percent of urban students in noncharter public schools and 74 percent of lottery applicants receive free or reduced-price lunch. Urban students and lottery applicants also have low average baseline scores: 0.43σ and 0.37σ below the state average in math and 0.43σ and 0.42σ below the average in English language arts (ELA). Regarding test scores, lottery applicants are representative of urban students overall.

In contrast, most students in nonurban areas are white: 84 percent of nonurban public school students and 88 percent of lottery applicants are white. Students in nonurban locations are of more affluent backgrounds and have better baseline academic outcomes. Twenty percent of public school students in nonurban areas and 13 percent of charter school applicants receive subsidized lunch. Nonurban students and lottery applicants score 0.15σ and 0.33σ above the state average in math and 0.16σ and 0.41σ above the average in ELA.

2.3 Empirical Framework

To estimate the impact of urban and nonurban charter schools on educational attainment and other outcomes, we take advantage of the natural experiment created by charter school lotteries. We use randomized lottery offers as instruments for charter school attendance at each type of charter school in a two-stage least squares (2SLS) strategy with multiple endogenous variables. We link charter school attendance to outcomes with an equation of the following form:

$$y_i = \sum_j \delta_j d_{ij} + X_i'\Gamma + \rho^u C_i^u + \rho^n C_i^n + \epsilon_i, \tag{2.1}$$

where y_i is an educational outcome for student i, such as degree attainment. Charter attendance is represented by type with C_i^u and C_i^n, which are indicators for attendance prior to when y_i occurs at an urban (u) or nonurban (n) charter school with a lottery. The effect of attending an urban or nonurban charter is captured by ρ^u and ρ^n, respectively. A vector of baseline characteristics, X_i, increases statistical precision and includes indicators for gender, race, special education, English learner status, and subsidized lunch status and a set of year of birth fixed effects. Key to our estimation strategy is the inclusion of "risk sets," indicated by d_{ij}, which are lottery fixed effects that account for the set of charter schools applied to by each student and include the application year and grade. The risks sets thus account for different probabilities of charter school attendance conditional on the number of schools applied to or a school's popularity. We use robust standard errors.

Randomized charter school lottery offers serve as instruments for charter school attendance, coded as mutually exclusive indicator variables: Z_{i1} represents an initial offer and Z_{i2} represents a waitlist offer. In a few cases, schools reported only initial or waitlist offer information; in such situations we include the school but only make use of the single source of offer variation. Thus, the first stage of our 2SLS framework is:

$$C_i^k = \sum_j \mu_j d_{ij} + X_i'\beta + \pi_1^u Z_{i1}^u + \pi_2^u Z_{i2}^u + \pi_1^n Z_{i1}^n + \pi_2^n Z_{i2}^n + \eta_i; k \in u, n, \tag{2.2}$$

where C_i^k indicates attendance at a charter school of k type, where $k \in u, n$, and is estimated as a function of the risk sets described above, the same vector of student characteristics, and the randomized lottery offers. The effect of lotteries on attendance is captured by π_1^k for the initial offer and π_2^k for the waitlist offer.

First stage estimates showing that charter offers boost charter attendance by 34 to 60 percentage points are presented in Appendix Table **??**. We demonstrate in Appendix Table B.2 that the characteristics of students offered seats in the lottery are very similar to not-offered students in both locations, offering a check on lottery randomization. Match rates to the SIMS data are above 99 percent and are very similar across lottery offers (Appendix Table B.3). We observe a small but statistically significant amount of differential attrition between lottery winners and losers for test scores and in 9th grade (Appendix Table **??**), prompting us to use Lee (2009) bounds for MCAS and high school outcomes (see Appendix A for a more detailed discussion). Our analysis is concerned primarily with college outcomes. For these outcomes there is no differential attrition and we have almost complete sample coverage (94 percent). We thus do not present bounds for college outcomes.

The control complier mean (CCM) is our preferred indicator for the counterfactual comparison (Katz et al., 2001; Abadie, 2002). The CCM is the average value of the outcome for compliers without charter school offers. These are students who do not attend a charter when they do not receive an initial or waitlist offer in the first charter school lottery they apply to. We estimate the CCM for each charter type k as follows (Katz et al., 2001; Abadie, 2002):

$$y_i * \left(1 - C_i^k\right) = \sum_j \lambda_j d_{ij} + X_i'\alpha + \tau\left(1 - C_i^k\right) + \nu_i; k \in u, n, \tag{2.3}$$

where τ is the estimate of the CCM and (1-C_i^k) is instrumented by the initial and waitlist offers, with risk sets and demographics accounted for as in Equation 2.

2.4 Results

2.4.1 Standardized Test Scores

MCAS math and ELA scores two years after the lottery serve as our benchmark to compare our findings to those of previous studies (Table 2.2). After two years, urban charters increase scores by almost half a standard deviation (σ) in math (0.48σ) and 0.32σ in ELA. These results align with the per-year effects found in APW of 0.33σ for middle-school math, 0.15σ for middle-school ELA, 0.34σ for high-school math, and 0.26σ for high-school ELA, though the comparison is inexact because of the different parameterizations. The urban results are also on par with those reported for Boston (Abdulkadiroğlu et al., 2011; Angrist et al., 2016; Walters, 2018; Cohodes et al., 2021; Setren, 2021; Cohodes and Feigenbaum, 2021).

After a student spends two years in a nonurban charter school, test scores drop by 0.11σ in math and 0.14σ in ELA. The corresponding per-year middle-school estimates from APW are -0.12σ for math and -0.14σ for middle-school ELA, with negative but not statistically significant impacts on high-school tests. Separating the sample into schools that exclusively serve middle-school grades (i.e. 6–8 or 5–8) and schools that offer high-school grades (e.g. 9–12 or 6–12) yields results that closely align with APW (Table 2.4).

We also present Lee bounds on the reduced form estimates to address differential attrition. Findings from this exercise suggest that, even in the presence of nonrandom attrition, the overarching test score patterns remain consistent with our main results. The bounds for MCAS scores in urban areas are very tight, given the minor differential attrition there. The nonurban upper bound is zero rather than negative, however, implying that if differential MCAS attrition is fully due to nonrandom selection into the sample, we would not find negative nonurban MCAS effects.

Notably, test score gains and losses occur at different points in the test score distribution. Comparison (traditional) urban students score approximately a third of a standard deviation below the state mean, whereas traditional nonurban students score almost half a standard deviation above the state mean. Thus, the test score gains in urban charters shift the distribution of scores rightward from below the state average to at or above the state average in two years, whereas nonurban charter students, despite their performance being lower than that of traditional nonurban students, still perform above the state mean (Appendix Figure B.1).

2.4.2 College Preparation

High-school students can prepare for college with several college-prep curricula, including AP courses, IB coursework, and other rigorous classes. We show the impact of charter school attendance on those outcomes in Table 2.2. Both urban and nonurban charters increase college-prep coursework, but via different paths.

In terms of AP preparation, in urban areas, charter attendance increases the AP-taking rate by 16 percentage points whereas in nonurban areas, charter attendance *decreases* AP test-taking by 29 percentage points. The decline is at least partly due to nonurban charters offering fewer APs. AP passing rates (scoring 3 or above) align with the change in AP-taking, with urban charters boosting scores of 3 or above by 5 percentage points and nonurban charters decreasing this rate by 19 percentage points (Appendix Table B.5). While AP courses are a popular college-readiness program, some high schools offer alternative paths. One nonurban charter school offers an IB curriculum; this results in nonurban charter attendance increasing the IB course-taking rate by 14 percentage points. We also examine enrollment in calculus regardless of AP Calculus enrollment, as it is an important college precursor. Calculus-taking rates increase for urban charter students by 5 percentage points and they decrease by a slightly smaller amount for nonurban charter students though this difference is not statistically significant.

AP, IB, and calculus may not encompass all college-ready curricular paths, especially in nonurban schools, many of which adhere to more individualized, project-based curricula. As a summation of college preparation, we turn to MassCore completion, which indicates a rigorous college-preparation curriculum as defined by the state. In urban areas, about 42 percent of comparison students meet the MassCore threshold; charter attendance increases this to 53 percent. In nonurban areas the charter bump is even larger: an increase from 74 percent MassCore completion for comparison students to 88 percent. Both urban and nonurban charter schools increase college preparation, with urban charters focusing on AP courses and MassCore and nonurban charters emphasizing the MassCore college-prep curriculum.

Taking, and scoring well on, the SAT test is another milestone on the path to college. As shown in Table 2.2, urban charter attendance increases SAT taking by about 4 percentage points, up from 63 percent for comparison students. Nonurban charter attendance does not change the rate of SAT-taking, with 79 percent of nonurban students in the sample taking the SAT. Urban charter attendance boosts the test scores of takers by 39 points (out of 1600), with little difference in nonurban scores.[6]

In order to matriculate to college, high-school students must also progress through high school and graduate. We display treatment estimates for high-school graduation in Table 2.2. Here, the findings diverge from those on test scores. Urban students are less likely to graduate high school on time, with a 7-percentage-point decrease in high-school graduation in four years. Urban charter students do catch up, with little difference in graduation rates vis-à-vis their peers' at the 5-year horizon. This is consistent with Angrist et al. (2016), which suggests that many Boston charter students take five years to graduate in order to complete high-school requirements, and our finding that urban charter attendance increases the likelihood that students repeat 9th or 10th grade by 4 percentage points (Appendix Table B.6). Nonurban charter students graduate at the same rates and within the same timeframe as their peers.

Since there is modest differential attrition in presence in high school in the nonurban areas, we

[6] We display SAT reasoning scores (out of 1600) since all cohorts take the relevant SAT subsections and only some take the exam scored out of 2400.

present Lee bounds for high-school outcomes as well. The bounds for these outcomes are quite tight and the small differences in appearance in the data do not affect our conclusions.

Overall, our findings suggest that attending an urban charter school boosts several measures of college preparation: students increase the number of APs taken, their completion of a college-ready high-school curriculum, the likelihood that they take the SAT, and their SAT scores. There are negative impacts on high-school graduation, which diminish over time. These estimates are similar to those previously reported for Boston charters (Angrist et al., 2016; Setren, 2021; Cohodes and Feigenbaum, 2021). For the first time, we present evidence on nonurban charter attendance on college preparation: nonurban charter attendees take fewer APs, given their schools' lower AP course offerings, but are much more likely to complete the rigorous MassCore curriculum. SAT and high-school graduation outcomes are unchanged by nonurban charter attendance.

2.4.3 College Enrollment

College preparation in high school is an important precursor to college, but college enrollment, persistence, and graduation show whether students succeed outside secondary education. Within a year of projected high-school graduation, both urban and nonurban charter students enroll in four-year college at greater rates than their peers, as shown in Figure 2.1 and Table 2.3. Additionally, both types divert enrollments from two-year institutions, such that initial enrollment in any post-secondary institution remains flat in both localities (Appendix Table B.7). Urban charter attendance boosts immediate four-year enrollment to 45 percent from 39 percent; nonurban charter attendance boosts enrollment to 62 percent from 53 percent (Figure 2.1). The decline in two-year college enrollment due to urban charter attendance is 4 percentage points and that due to nonurban charter attendance is 7 percentage points. By the second year after projected high-school graduation, an interval that allows for late high-school graduation, there is little difference in enrollment at two-year institutions, and four-year college enrollment increases by 8 and 10 percentage points for urban and nonurban charters, respectively. Since two-year enrollment changes little and four-year enrollment rises, enrollment in any college increases for both charter types in the second year after projected high-school graduation, as shown in Table 2.3. With no differential attrition in the college data, we do not present bounds for these or other college outcomes.

Following the time trend in Figure 2.1 into the 3rd and 4th years after expected high-school graduation, urban charters boost four-year college enrollment by 5 to 7 percentage points, with the control complier mean decreasing over time as students drop out. In the 5th and 6th years after high-school graduation, urban charters increase enrollment, though the interpretation of this outcome is ambiguous: If it represents progress toward a degree, enrollment could be beneficial; if it represents a delay in joining the workforce, it could be detrimental. The decrease in control complier enrollment is now due in part to graduation from college. Nonurban charters boost four-year enrollment in the 3rd and 4th years by 10 to 13 percentage points, with lower dropout among the counterfactual students. Nonurban charter students are also more likely to be enrolled in the 5th and 6th years after projected high-school graduation by 4 to 5 percentage points, though only the 6th year difference is statistically significant. Urban and nonurban charters increase both initial college enrollment and persistence through college.

2.4.4 Degree Attainment

Both urban and nonurban charter school attendance increases the likelihood that a student obtains any degree, in particular a bachelor's from a four-year institution. In the 4th year after projected high-school graduation, which corresponds to on-time high-school progress and on-time college progress, urban charters increase BA receipt by 3 percentage points and nonurban charters by 7 percentage points (Figure 2.1). Urban charters boost two-year attainment by a small amount, whereas nonurban schools decrease it, meaning that both school types increase receipt of a degree of any type by 4 to 6 percentage points (Appendix Table B.8).

As time goes on, urban charter attendance increases the BA boost to 4.2 percentage points by the 6th year after projected high-school graduation and the gains in any degree attainment to 4.6 percentage points. The nonurban edge increases to an even greater extent over time, with a bump of 11.9 percentage points for BA attainment (11.2 percentage points for any degree) in the 6th year after projected high-school graduation (Table 2.3). By the 6th year, 22 percent of urban control compliers graduate with a BA, with urban charter attendance increasing this to 26 percent, an increase of 19 percent of the comparison mean. By the 6th year, 47 percent of the nonurban control compliers graduate college, with the charter effect boosting this outcome for treated compliers to 59 percent, a 25 percent increase over the mean. Charter attendees are more likely to enroll and graduate from four-year colleges in both urban and nonurban areas. Our conclusions are robust to excluding covariates, adding baseline scores, or instrumenting using initial offers only (Appendix Table B.9). The nonurban charter effect is even larger for college graduation outcomes than for college enrollment outcomes. Notably, 6 years after high-school graduation, the college graduation edge from nonurban charter attendance is more than twice as large as that from urban charter attendance.

2.4.5 College Quality

College quality can increase college graduation and earnings (Hoekstra, 2009; DeAngelo et al., 2011; Cohodes and Goodman, 2014; Zimmerman, 2014; Goodman et al., 2017; Ge et al., 2022; Black et al., 2023). Thus, we investigate the impact of charter attendance on college quality using college categories from the Barron's Guide and college characteristics from the US Department of Education's Integrated Postsecondary Education Data System (IPEDS) in Table 2.3. IPEDS college information is assigned to students by their first enrolled institution. We also consider the extent to which college quality accounts for the observed boost in graduation. Urban charter attendance increases both four-year college enrollment and BA attainment in fairly equal measure at highly competitive institutions and competitive institutions. Nonurban charter attendance boosts college enrollment primarily at highly competitive institutions and graduation at highly competitive and competitive institutions. The differences in college atmospheres are reflected in increased instructional expenditures per student (compensation for instruction divided by number of students) of \$809 for urban college attendees and \$928 (not significant) for nonurban. Nonurban college students attend institutions with lower student/faculty ratios (number of non-graduate students divided by number of non-graduate-teaching instructional staff).

To measure institution graduation rate, we use the IPEDS' 150% completion rate, which reflects graduation within 6 years for 4-year programs and within 3 years for 2-year programs. Urban charter attendance improves the institutional graduation rate by 4.3 percentage points. Nonurban charter attendance also boosts the graduation rate of the institution attended by 4.3 percentage points. The graduation rate shift due to urban charter attendance almost exactly matches the boost in any degree attainment (4.6 percentage points) whereas in nonurban areas, the shift in graduation rates is about 40 percent of the change in degree attainment ($\frac{0.043}{0.112}$). The shift in college quality appears to explain the degree gains for urban charters but not for nonurban charters.

We can show this another way by considering the implicit 6-year graduation rates for treated and untreated compliers at 4-year colleges. In urban areas, the graduation rate for comparison students is 58.9 percent ($\frac{0.216}{0.367}$). For treated students, the graduation rate is a similar 57.8 percent ($\frac{0.216+0.042}{0.367+0.079}$). In nonurban areas, the graduation rate is 85 percent for counterfactual students ($\frac{0.472}{0.557}$), whereas it is 88.9 percent for treated students ($\frac{0.472+0.119}{0.557+0.098}$). This implies that the nonurban charter college boost operates not only through enrolling in high quality institutions but also through how students experience and complete college.

2.4.6 Alternative School Groupings

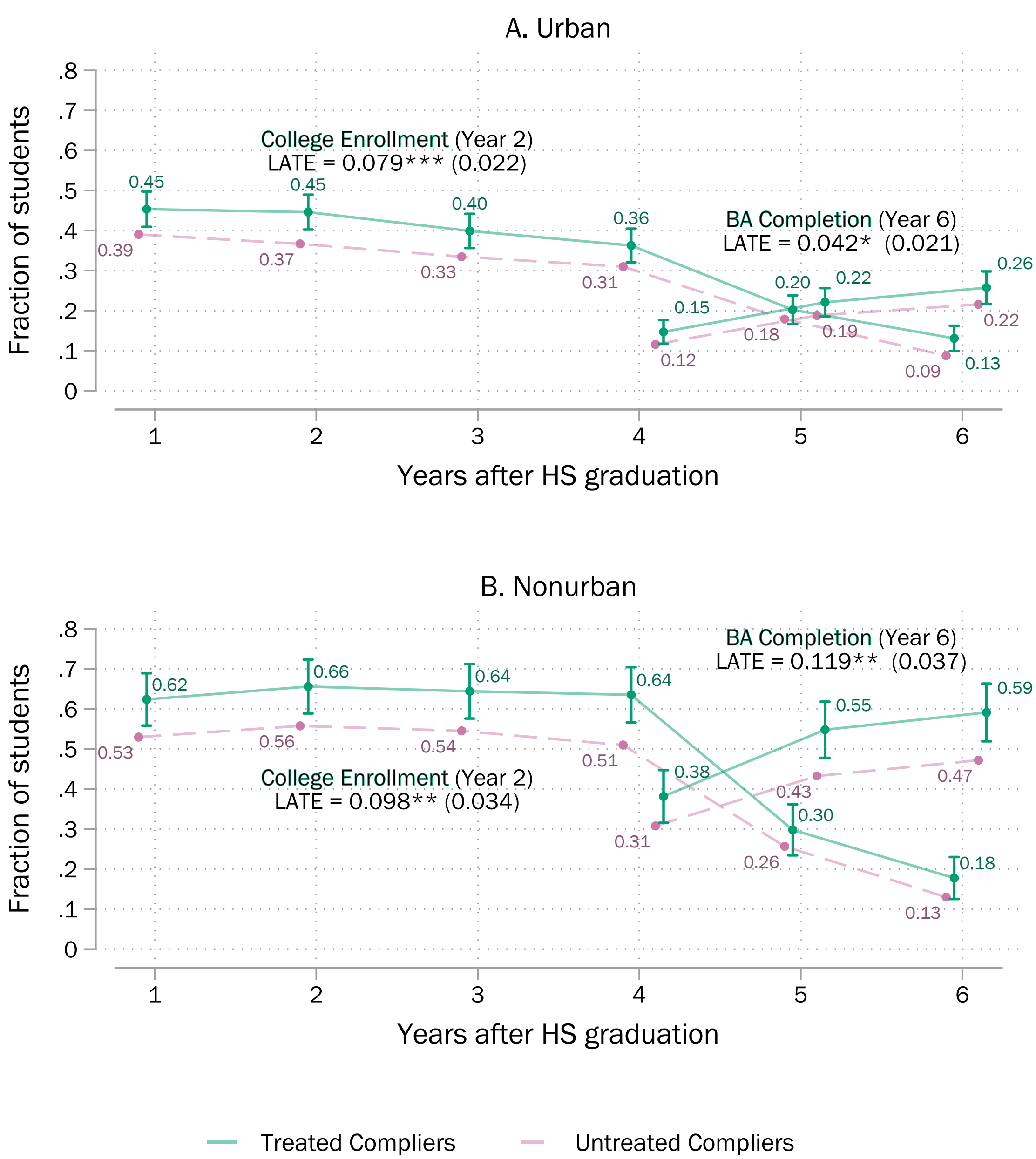
In Table 2.4, we present results with alternative school groupings. The college gains for both localities are concentrated among schools that offer high-school grades (a majority of our sample), with no and perhaps negative effects on college for the few schools that only offer middle-school grades, similar to the evidence from Place and Gleason (2019) and Demers et al. (2017). This finding indicates that continuity between the charter school environment and college transition may be a key factor behind the college boost. We also regroup schools by their practices rather than their localities. One nonurban school in our sample follows No Excuses practices and two urban schools do not follow the No Excuses model. Under this categorization the No Excuses gains are slightly larger than the urban ones, and the non-No Excuses schools generally follow the pattern for the nonurban schools. However, the test score results are null rather than negative and the college boost is slightly smaller. This analysis reinforces the notion that multiple school models can lift college outcomes and that test score gains are not a necessary precursor to college gains.

2.5 Conclusion

We confirm previous evidence from Massachusetts that urban charters boost test scores, whereas nonurban charters do not, a pattern that aligns with results in the broader charter school literature. However, when we turn to college enrollment and graduation, we have several novel findings. First, we show that the bump in college enrollment found previously for Boston charter attendance translates into degree completion in a wider sample of urban schools, with urban charters boosting BA attainment rates by 4.2 percentage points and attainment of any degree by 4.6 percentage points within 6 years. Second, we show that nonurban charter schools—the same schools that do not increase test scores—increase four-year college enrollment and BA attainment by 9.8 and 11.9 percentage points, respectively. The current analysis cannot speak to all of the mechanisms behind the college gains, but we present evidence on a few key factors. Offering high-school grade levels seems to be a necessary condition for college impact, with no college gains for schools that offer only middle-school grades. Charters in both locales boost college-ready curricula, via AP and MassCore in urban areas and MassCore in nonurban ones. This academic preparation may be a key factor for the persistence effects. The college institutions that students enroll in matters: The urban charter college edge exactly parallels the rise in graduation rates due to type of enrolled institution, whereas for nonurban charters it accounts for 40 percent of gains. The remaining nonurban advantage is unexplained. In future work, we will investigate more of the mechanisms behind this pattern of results, including differences in school practices and contexts.

We draw two main conclusions from these findings. First, multiple charter school models can induce college gains. While many have focused on the "No Excuses" practices as key to charter school success, the nonurban schools in this sample operating on alternative models deliver a large boost to BA attainment. Second, although test scores and longer-term outcomes are typically positively correlated, we add to the evidence that shows that the relationship between test scores and college outcomes does not hold in all contexts, concluding that researchers and policymakers should be wary of evaluating programs solely on standardized test results.

Figure 2.1: Four-Year College Progression



Notes: This figure shows the treatment and control complier means for four-year college enrollment and graduation for treated and untreated compliers. Treatment effects on college enrollment two years after projected high-school graduation and graduation rates six years after projected high-school graduation are reported under the labels. Details for the other treatment effects are in Appendix Tables B.7 and B.8. Robust standard errors in parentheses (+ p<0.10 * p<0.05 ** p<0.01 ***p<0.001).

Table 2.1: School and Student Characteristics

	Charter Schools		Other Public Schools	
	Urban (1)	Nonurban (2)	Urban (3)	Nonurban (4)
(A) Schools: Administrative Records				
% of teachers licensed in subject	58.867	70.611	96.608	96.855
% of core classes taught by highly qualified teachers	86.227	94.089	89.876	97.378
Title 1 school	1.000	0.667	0.767	0.409
Student-teacher ratio	11.721	10.511	13.955	13.165
Per-pupil expenditure	16250	11982	15661	14411
School size	433	435	663	2271
Counselors per 1000 students	5.204	2.150	2.793	3.043
Disciplined students per 1000 students	191.923	33.963	122.090	45.412
(B) Schools: Survey Responses				
Days per school year	192	182	-	-
Hours per school day	7.935	6.974	-	-
High-quality tutoring	0.615	0.111	-	-
Frequent teacher observations	0.538	0.375	-	-
Frequent checks for student understanding	0.846	0.500	-	-
Differentiated instruction	0.692	0.750	-	-
Culture of high expectations	0.733	0.111	-	-
Project-based learning	0.231	0.625	-	-
N(Schools)	15	9	266	599
(C) Students: Baseline Characteristics				
Female	0.518	0.519	0.483	0.491
Asian	0.030	0.030	0.076	0.039
Black	0.526	0.026	0.199	0.046
Latinx	0.284	0.036	0.316	0.062
Other race	0.041	0.028	0.028	0.017
White	0.119	0.880	0.380	0.836
Special education	0.194	0.159	0.187	0.177
English learner	0.114	0.011	0.179	0.028
Free/reduced price lunch	0.740	0.128	0.644	0.198
Baseline MCAS ELA	-0.418	0.417	-0.432	0.155
Baseline MCAS Math	-0.364	0.331	-0.426	0.153
N	14,191	3,583	276,401	643,585

Notes: This table shows characteristics for urban and nonurban charter schools in the lottery analysis sample and lottery applicants in Columns 1 and 2. Information on traditional public schools that serve 6th and/or 9th grades in urban and nonurban areas and their students appears in Columns 3 and 4 for comparative purposes. Data sources for Panel A are Massachusetts Department of Elementary and Secondary Education School District Profiles for the 2013–2014 school year. Title I eligibility is reported for the 2013–2014 school year and comes from the U.S. Department of Education Common Core of Data (CCD). The data for Panel B come from a survey of charter school leaders fielded in 2011 and 2012. The survey response rate was 87.5% (12 out of 15 urban schools, all 9 nonurban schools). Panel C uses the student-level data for charter school applicants enrolled in schools in the state of Massachusetts at the time of application in the projected high-school classes of 2006–2018 in Columns 1 and 2 and for students who attended schools in the state of Massachusetts in 9th grade in the projected high-school classes of 2006–2018 in Columns 3 and 4.

Table 2.2: The Impact of Charter School Attendance on Tests and High School Academics

	MCAS		HS Curriculum				SAT		HS Graduation	
	Math	ELA	Took AP	Took IB	Took Calc	MassCore	Took SAT	Score	4 Year	5 Year
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(A) Urban										
2SLS	0.475***	0.318***	0.164***	-0.018**	0.052**	0.105***	0.042*	39.026***	-0.066***	-0.018
	(0.040)	(0.038)	(0.023)	(0.009)	(0.025)	(0.025)	(0.023)	(9.459)	(0.023)	(0.022)
CCM	-0.299	-0.332	0.323	0.023	0.166	0.423	0.626	892.771	0.655	0.733
N	10,706	10,822	10,831	5,938	5,938	9,761	10,831	6,951	11,983	11,983
Reduced Form	0.226***	0.151***	0.076***	-0.009*	0.029**	0.046***	0.014	17.984***	-0.029***	-0.009
	(0.020)	(0.019)	(0.011)	(0.005)	(0.013)	(0.012)	(0.011)	(4.512)	(0.011)	(0.010)
Lower Bound	0.192***	0.114***	0.076***	-0.009*	0.029**	0.046***	0.014	16.577***	-0.030***	-0.011
	(0.020)	(0.019)	(0.011)	(0.005)	(0.013)	(0.012)	(0.011)	(4.486)	(0.011)	(0.010)
Upper Bound	0.271***	0.215***	0.076***	-0.009*	0.031**	0.046***	0.014	18.319***	-0.028***	-0.008
	(0.019)	(0.018)	(0.011)	(0.005)	(0.013)	(0.012)	(0.011)	(4.512)	(0.011)	(0.010)
(B) Nonurban										
2SLS	-0.114*	-0.139***	-0.289***	0.142***	-0.037	0.140***	0.010	11.389	-0.016	-0.012
	(0.059)	(0.051)	(0.035)	(0.031)	(0.047)	(0.033)	(0.032)	(15.316)	(0.026)	(0.024)
CCM	0.372	0.481	0.512	0.066	0.289	0.737	0.786	1121.508	0.815	0.903
N	3,198	3,164	3,081	1,729	1,729	2,942	3,081	2,376	3,175	3,175
Reduced Form	-0.069**	-0.082***	-0.159***	0.078***	-0.014	0.083***	0.012	6.982	-0.013	-0.011
	(0.035)	(0.029)	(0.021)	(0.017)	(0.027)	(0.019)	(0.018)	(8.987)	(0.015)	(0.014)
Lower Bound	-0.113***	-0.120***	-0.165***	0.075***	-0.016	0.077***	0.005	-5.198	-0.018	-0.015
	(0.035)	(0.029)	(0.020)	(0.017)	(0.027)	(0.019)	(0.018)	(8.819)	(0.015)	(0.014)
Upper Bound	0.012	0.004	-0.153***	0.080***	-0.012	0.090***	0.017	18.507**	-0.009	-0.005
	(0.034)	(0.027)	(0.021)	(0.017)	(0.027)	(0.019)	(0.018)	(8.880)	(0.015)	(0.014)

Notes: Each coefficient labeled 2SLS is the instrumental variables estimate of the effect of attending an urban or nonurban charter on the outcome listed in the column heading as described in Equation 2.1. The control complier mean is listed in the rows labeled CCM. The estimate of the effect of winning an urban or nonurban charter lottery is reported in the rows labeled Reduced Form. For concision, the reduced form estimates use a single offer variable, ever offer, in each locality. Estimates of the reduced form using Lee Bounds are labeled Lower Bound and Upper Bound. The sample includes charter lottery applicants in the projected high-school classes of 2006–2018. MassCore refers to graduating within 5 years with a college-preperation curriculum as defined by the state. It is available for the classes of 2008–2018. The IB and Calculus indicators are available for the classes of 2015–2018. Robust standard errors in parentheses (+ p<0.10 * p<0.05 ** p<0.01 ***p<0.001).

Table 2.3: The Impact of Charter School Attendance on College Outcomes

	Urban			Nonurban		
	2SLS	CCM	N	2SLS	CCM	N
	(1)	(2)	(3)	(4)	(5)	(6)
(A) College Enrollment (in Y2)						
All	0.077***	0.495	13,281	0.063+	0.684	3,414
	(0.023)			(0.032)		
All: Graduation rate (IPEDS)	0.043**	0.447	7,993	0.043*	0.573	2,728
	(0.014)			(0.019)		
All: Instructional exp./student	809.324*	7184.252	7,849	927.703	9258.788	2,711
	(332.370)			(757.111)		
All: Student/faculty ratio	-0.061	15.748	7,993	-0.530+	14.921	2,728
	(0.220)			(0.310)		
2 Year	-0.002	0.127	13,281	-0.035	0.127	3,414
	(0.016)			(0.023)		
4 Year	0.079***	0.367	13,281	0.098**	0.557	3,414
	(0.022)			(0.034)		
4 Year: Highly Competitive	0.036*	0.118	13,281	0.065+	0.317	3,414
	(0.016)			(0.034)		
4 Year: Competitive	0.044*	0.185	13,281	0.036	0.204	3,414
	(0.019)			(0.031)		
4 Year: Noncompetitive	0.001	0.062	13,281	-0.001	0.036	3,414
	(0.011)			(0.015)		
(B) College Degrees (by Y6)						
All	0.046*	0.240	11,608	0.112**	0.519	3,158
	(0.022)			(0.037)		
AA	0.009	0.041	11,608	-0.030	0.093	3,158
	(0.011)			(0.020)		
BA	0.042*	0.216	11,608	0.119**	0.472	3,158
	(0.021)			(0.037)		
BA: Highly Competitive	0.023+	0.082	11,608	0.078*	0.260	3,158
	(0.014)			(0.034)		
BA: Competitive	0.022	0.095	11,608	0.052+	0.161	3,158
	(0.015)			(0.029)		
BA: Noncompetitive	-0.003	0.037	11,608	-0.011	0.049	3,158
	(0.010)			(0.016)		

Notes: Each coefficient in columns labeled 2SLS is the instrumental variables estimate of the effect of attending an urban or nonurban charter on the outcome listed in the column heading as described in Equation 2.1. The control complier mean is listed in the column labeled CCM. The sample includes charter lottery applicants in the projected high-school classes of 2006–2018. Highly Competitive includes Barron's categories *highly competitive*, *most competitive*, and *very competitive*; Competitive includes the categories *competitive* and *special*; and Noncompetitive includes *noncompetitive*, *unranked*, and *less competitive*. College characteristics come from IPEDS and are availabe for students enrolled in college: graduation rate is the 150% graduation rate; instructional expenditures/student is the institution-reported expenditures on instructional staff compensation divided by the number of full-time equivalent students; and student/faculty ratio is the number of full-time equivalent non-graduate students divided by the number of full-time equivalent instructional staff teaching non-graduate students. Robust standard errors in parentheses (+ $p<0.10$ * $p<0.05$ ** $p<0.01$ *** $p<0.001$).

Table 2.4: The Impact of Charter School Attendance on Key Outcomes, Alternative Samples

	Math MCAS	4-Year College Enrollment Urban	4-Year College Graduation	Math MCAS	4-year College Enrollment Nonurban	4-year College Graduation
	(1)	(2)	(3)	(4)	(5)	(6)
Main specification	0.475***	0.079***	0.042*	-0.114+	0.098**	0.119**
	(0.040)	(0.022)	(0.021)	(0.059)	(0.034)	(0.037)
N	10,706	13,281	11,608	3,198	3,414	3,158
Serves HS grades	0.418***	0.095***	0.062**	-0.039	0.125**	0.153***
	(0.044)	(0.025)	(0.023)	(0.071)	(0.041)	(0.043)
N	9,001	11,294	10,236	2,494	2,689	2,475
Serves MS only	0.608***	0.011	-0.072+	-0.346***	0.020	0.010
	(0.071)	(0.040)	(0.038)	(0.097)	(0.063)	(0.067)
N	1,690	1,983	1,351	704	725	683
		No Excuses			Non-No Excuses	
	(7)	(8)	(9)	(10)	(11)	(12)
Alt. Grouping	0.494***	0.078***	0.043*	-0.021	0.070*	0.075*
	(0.038)	(0.022)	(0.020)	(0.058)	(0.033)	(0.033)
N	10,071	12,402	10,821	4,937	5,949	5,367

Notes: The first row of the table repeats the main specification reported in the other tables, see Tables 2.2 and 2.3 for details. Four-year college enrollment is enrollment in 2 years of projected high school graduation. Each subsequent row shows an alternative sample definition. The row labeled Serves HS grades restricts the sample to schools that offer high-school grade levels. The row labeled Serves MS only restricts the sample to schools that exclusively offer middle-school grade levels and do not offer high-school grade levels.The row labled Alt. Grouping groups schools by No Excuses and not No Excuses instead of urban and nonurban. Since students may apply to both types of schools, they may be counted in the N for each type. Robust standard errors in parentheses (+ p<0.10 * p<0.05 ** p<0.01 ***p<0.001).

Chapter 3. Learning to Read in Fragile Contexts: Experimental Evidence from a Comprehensive Literacy Intervention in Haiti[1]

3.1 Introduction

Despite increasing levels of enrollment, 70 percent of children in low and middle income countries are not able to read and understand a simple story by age 10 (World Bank, 2022). Globally, over 500 million children were failing to meet minimum levels of proficiency in reading by the end of 2020 (UNESCO, 2021). Early literacy skills are arguably the most important skills for children's learning experience, as the acquisition of other skills, such as numeracy, heavily depends on children's reading comprehension (Crawford et al., 2020). Moreover, these types of skills have been shown to be an important determinant of children's long-term academic achievement (Duncan et al., 2007; Stern et al., 2023) and labor market outcomes, including earnings and employment (Hanushek et al., 2015; McIntosh and Vignoles, 2001). For these reasons, improving literacy is a major focus of current efforts in the international community. Understanding how to improve skill acquisition among young children, particularly in contexts with limited implementation capacity, is important for developing effective education policies in the developing world.

Existing literature demonstrates the effectiveness of multi-pronged, or 'bundled,' approaches in improving learning outcomes in countries with extremely low learning levels (Fazzio et al., 2020; Eble et al., 2021; Banerjee et al., 2015; Piper et al., 2018a; Borman et al., 2007). For instance,

[1]Co-authored with Melissa Adelman, Juan Baron, José Mola, and Atabanam Simbou.

Piper et al. (2018a) show that combining multiple elements—such as teacher training, coaching, and high-quality books—into a coherent program is considerably more effective than stand-alone interventions. However, these types of multi-pronged interventions have often taken place outside of the regular school system, or have relied heavily on external monitoring, limiting their ability to be replicated at a larger scale or to be continued in the medium to long term.

In this paper, we look at the impact of M'ap Li Net Ale ("I am reading all the way"), a comprehensive early-grade reading intervention including (i) training of teachers and school principals, (ii) provision of materials including textbooks and teacher guides with structured lesson plans, and (iii) periodic coaching by an external team or by school principals. The intervention, implemented in Haiti, also included activities for community mobilization and extracurricular classes for students. By simultaneously tackling various inputs of the learning process, the intervention is close in nature to the Success for All whole-school reform model implemented in the United States, which similarly focused on early literacy and provided schools with training, coaching support, and learning materials (What Works Clearinghouse, 2009).[2]

In Haiti, the context of our study, a version of this intervention was implemented and evaluated in 2012-2014 (RTI International, 2015). An evaluation report found positive effects on literacy outcomes, but highlighted the challenges of providing effective external coaching, including the costs and logistical difficulties of regularly visiting schools (RTI International, 2015).[3] Prompted by the search for more scalable approaches to improve literacy, in this study we test two versions of this intervention: (i) a "Full" approach relying on an external team for coaching and (ii) a "Light" approach where coaching was fully led by school principals, and which was approximately 15 percent less costly. We find that 18 months after the intervention started, the less costly approach

[2]Extensive evaluations of the Success for All model have been conducted in the United States, as documented by Borman et al. (2007, 2005a,b). See Borman et al. (2002) for a meta-analysis.

[3]RTI and USAID later implemented a revised version in 2014-2016, in parallel to the intervention evaluated here, but in a different set of schools. An evaluation of this version found positive effects in 6 of 8 reading skills measured (Guzmán et al., 2021).

relying on school principals was as effective as the Full approach across all dimensions of learning assessed by Early Grade Reading Assessments (EGRA). Both versions of the intervention increased overall EGRA scores by 25 to 27 percent (XX SD).

Contrary to the idea that radical or "takeover" approaches are needed to achieve large gains in learning outcomes in the developing world, our results suggest that interventions operationalized within existing, albeit weak, institutional frameworks can not only be effective in improving learning outcomes, but can be as effective as other approaches relying on external teams, at substantially lower costs. The findings presented here point out a scalable approach to improve reading in fragile settings, leveraging school capacity to achieve sustainable, large-scale improvements in learning outcomes.

3.2 Background and Intervention Design

With an estimated income per capita of USD 1150, Haiti is the poorest country in the Western Hemisphere and ranks 170th out of 189 countries in the 2020 Human Development Index (HDI), a multi-dimensional measure of wellbeing (World Bank, 2021). Political instability, high levels of insecurity, and frequent natural disasters undermine economic and social development efforts and contribute to the limited public provision of basic services, including sanitation, education, and healthcare. In some cases, local and international actors have stepped in to fill the void: more than 80 percent of primary schools in Haiti are privately owned and operated by a variety of actors including religious networks, non-governmental organizations, and private citizens (Baron et al., 2016). This private provision, partially supported by donor and national financing, has helped increase the average years of completed schooling to 11 years today (World Bank, 2020). However, the challenging socioeconomic context facing most families and the low average quality of education provided in schools means that 55 percent of 10-year-olds cannot read and understand a simple story by the end of primary school.

In schools across the globe, teachers are considered the most important input for learning, and previous research in Haiti suggests that teachers' methods are an important contributor to low learning outcomes (Bruns and Luque, 2014; Adelman et al., 2015; World Bank, 2017b). Specifically, Adelman et al. (2015) find that although primary teachers are present in their classrooms and teaching, they rely on ineffective methods (like lecturing and unison response) and are much more likely to teach reading and writing in French, rather than Haitian Creole, most students' mother tongue and the first language that should be introduced according to official education policy (DeGraff, 2016).

In this paper, we evaluate M'ap Li Net Ale, a comprehensive literacy program implemented between 2014 and 2016 that targeted the early grades of primary school. Our sample is composed of a total of 446 public schools. These schools are geographically dispersed, covering four of the countries' ten departments (Grande-Anse, Nippes, Ouest, and Nord-Ouest).[4] The two versions of the M'ap Lit Net Ale program evaluated in this paper were developed over successive iterations supported by the World Bank and the United States Agency for International Development (USAID). The Ministry of National Education and Vocational Training (MENFP) and the Episcopal Commission for Catholic Education (CEEC) led its implementation. The program aimed to improve early-grade literacy by focusing on the most important components of good reading acquisition in terms of ability, speed, and fluency in the early years of basic education. The design of the program reflects at least three elements that have emerged in recent years as best practice for improving early-grade literacy: i) mother tongue instruction, ii) the use of ongoing coaching sessions, and iii) providing structured quality teaching-learning materials along with training and coaching.

As mentioned above, most schools in Haiti teach in French, the second official language, even though only 5 percent of the population speaks French fluently. Haitian Creole, on the other hand, is spoken by all children in Haiti. Despite this, the latter language has been largely excluded from the school system. Considering evidence showing that children taught in their own language are

[4]See Table 3.1 in the Appendix for the geographic distribution of schools and students in our sample.

more likely to enroll and succeed in school (Kosonen, 2005; World Bank, 2021), the program's instruction materials and accompanying textbooks were developed in Creole.

The M'ap Li Net Ale program's structure reflects emerging evidence on effective strategies for literacy improvement. Piper et al. (2018b) highlight the importance of structured approaches to pedagogy and the institutional infrastructure needed for their success. In particular, they emphasize the need to accompany teacher support in the form of coaching with high-quality learning materials. The program studied in this paper aimed to combine teacher training and coaching with the provision of teaching guides (including structured lesson plans) and student textbooks.

The intervention consisted of three main components: training, coaching, and the distribution of school materials. The first two components aimed to strengthen teachers' skills for teaching reading and writing through training and regular coaching. At the onset and during the course of the intervention, teachers and principals participated in a 5 to 8-day general training covering two main stages: formation (Full and Light treatment) and remediation (Full treatment only). As part of this training, a group of external coaches and school principals were taught to supervise and support teachers in participating schools through "coaching sessions," which took place two times per month throughout the second semester of the school year. During these sessions, coaches supported teachers in adopting remedial measures and providing reinforcement to students.

The third main component intended to address current gaps in school supplies both at the teacher and student level, by providing teachers with teaching resources, including classroom posters and teaching guides with scripted lesson plans, and by providing students with textbooks and kits with school materials. Alongside these three core components, the program offered further support through small libraries and promoted community engagement through assistance for School Councils (SC) and Parents' Committees (PC). For example, classrooms received small libraries containing titles in French and Creole, and teachers were encouraged to organize book clubs, to provide time for storybook reading during the school day, and to hold reading competitions to boost the practice of reading among students. SC members were trained to develop plans for school development and improvement (PDAS), and to set up and manage small projects for their respective schools, with the goal of strengthening school governance. After the training, SC members were to develop their own PDAS as small projects for improving the school environment. Periodic supervision was provided by an external team.

Overall, the activities were very similar in the Full and Light versions of the program. Both interventions included training and regular coaching for teachers and principals, provision of teaching supplies and school materials, general supervision by an external team, and community mobilization activities. The main differences between the interventions had to do with coaching activities and the length and content of training sessions. In the Full version, a team of external coaches led all teacher coaching activities, whereas in the Light version it was principals who led this process. Moreover, in the Full version, principals and teachers were trained for a total of 8 days (8 hours per day); in the Light version they were trained for 5 days. The three additional training days provided in the Full version covered remediation topics and were not offered to principals in the Light version. These differences accounted for a 19 percent reduction in implementation costs for the Light program during the first year, and 7 percent during the second.

A previous evaluation of the Full program found it effective in improving literacy but highlighted challenges related to providing external coaching, such as the high costs and logistical difficulties of regular school visits (RTI International, 2015).[5] The RTI evaluation aimed to compare the effectiveness of a close variant of the Full program with (Treatment B) and without (Treatment A) community-based activities, against a control group. The report showed gains in both treatment groups during the 2013–2014 school year in initial sound identification and letter sound knowledge, with larger gains for Treatment A, but no improvements in reading comprehension. Two subsequent evaluations, covering 2014-2016 and 2018-2020, followed this one (Guzmán et al., 2021;

[5]The program evaluated here is an adapted version of "Tout Timoun Ap Li" (All Children Reading). This literacy intervention was led by USAID between 2012 and 2016.

D'Agostino et al., 2024). The first found positive effects at the end of grades 1 and 2 for most EGRA subtasks except for reading and listening comprehension. The second evaluation found significant improvements in seven out of nine subtasks for a cohort that received the treatment for two years, including gains in reading comprehension.[6]

The high costs and implementation challenges highlighted in previous evaluations motivated the search for a less costly and more administratively feasible approach to implement the program. The Light version of M'ap Li Net Ale was developed and tested in an effort to identify a version of the program that could feasibly be implemented at a national scale in the challenging context of Haiti while maintaining its effectiveness.

In both versions of the program, principals played a key role in its implementation. They were responsible for supervising teachers' classroom work and ensuring adherence to the program's practices. To effectively fulfill this role, principals were required to attend all training sessions and become well-versed in the M'ap Li Net Ale framework and its practical application. Additionally, they were expected to intervene and provide guidance when teachers deviated from the established practices. The primary goal of the intervention was to build the capacity of schools to independently apply the method, reducing reliance on external support. Therefore, training both principals and teachers was essential.

The implementation of M'ap Li Net Ale covered two successive school years for students who attended 1st grade in the 2014-15 and 2015-16 school years. The first phase of the implementation took place during the 2014-2015 school year and the second phase during 2015-2016. In both phases, the program was implemented from January to May, the second half of the school year.[7]

[6]All three evaluations correspond to "Tout Timoun Ap Li" or iterations of this program: "Read to Learn" (2014-216) and "Read Haiti" (2016-2020). These interventions were led by USAID and were implemented in a sample of schools different from ours.

[7]A full timeline including more detail on the intervention components can be found in Appendix Figure C.2.

3.3 Data

We conducted three waves of data collection: (i) a baseline survey at the beginning of February 2015 covering the cohort of students attending the 1st grade during this year, (ii) a midline survey in November 2015, and (iii) an endline survey in May 2016 covering both cohorts (students who attended 1st grade in 2014-15 and continued to 2015-16 and students who started 1st grade in 2015-16).

At baseline, we collected data from 1st grade students, 1st grade teachers, and school principals. Midline and endline surveys were administered to the first cohort of students, now enrolled in the 2nd grade, and to a new cohort of students attending the 1st grade. Among students, our surveys collected data on students' reading skills and families' socioeconomic conditions. Teachers' surveys gathered information on teachers' characteristics, and information on pedagogy and classroom management. Teachers' surveys also collected data on student attendance and enrollment status. Lastly, principals were surveyed on basic school physical characteristics (equipment, infrastructure, etc.), school environment, school management, and characteristics of school principals. In midline and endline surveys, we focused on administering learning assessments. Unfortunately, this data is not longitudinal and we are unable to track students over time. Moreover, because of the large sample size (446 schools) and the complex logistics of administering learning assessments, random representative sub-samples of students were chosen at each evaluation period to take the tests. A timeline of the intervention and data collection efforts is shown in Figure 3.1.

3.4 Research Design

This research aims to answer two main questions regarding the *M'ap Li Net Ale* program. First, what is the impact of the program on students' learning outcomes? Second, can this impact be achieved at a lower cost (i.e., by making principals responsible for coaching rather than external staff)? Our study aims to answer these questions through a randomized controlled trial covering

446 schools in Haiti.

3.4.1 Sample

The design of this evaluation is based on school-level randomization. One condition defined eligibility for schools participating in this program, namely: to be a school that had received school grants as part of the government's Tuition Waiver Program, a World Bank-funded program that provided an annual per-student payment to private schools to cover students' tuition fees. After identifying all eligible schools, 446 were randomly selected to constitute the evaluation sample. Table C.4 in the Appendix presents baseline characteristics for the full sample selected for the study. This table provides a summary of the difficult socioeconomic conditions facing students.

About one in four pupils does not have a notebook and a pencil in their bag on the day of the visit. Only 33 percent of students say they have electricity at home, less than 50 percent have a radio, and about 20 percent own a television. Only 12 percent of the students' parents own a refrigerator, a bicycle or a motorbike, and less than 6 percent own a computer or a car.

We stratified the sample by department to ensure that each of the four departments is proportionately represented in each treatment or control group. Table 3.1 contains the number of schools and students by treatment group and geographic department at baseline (see Annex C Table C.5 for midline and endline information). An average of 16 students per school were selected randomly for baseline data collection and 29 per school for midline and endline assessments.

Once the evaluation sample was selected, the 446 schools were divided into three groups (Control, Light, and Full) through simple random assignment. By randomizing, we ensure that the three groups are on average statistically similar at baseline. In Table 3.2, we provide summary statistics of students' socioeconomic characteristics at baseline. Overall, the characteristics of the three groups of students are statistically balanced. Among twenty socioeconomic variables included in our survey, only three show small statistically significant deviations between groups (i.e., having a pencil, a notebook, and a computer). These differences seem to favor the control group.

In Appendix C, we present a baseline comparison of schools' characteristics across the three groups (see Table C.8 and Table C.9). We also present descriptive statistics for the full sample of schools. Our data shows that only 67 percent of schools offer pre-school education and that more than 70 percent of schools do not have third-cycle grades (grades 7 through 9). For grades 1 and 2, schools have an average enrollment of 32 students. Most schools offer single-shift lessons (88 percent), with an average duration of five teaching hours. We do not find statistically significant differences between the three groups across these variables. Some schools in our sample reported having participated on other programs such as the school canteen program (23 percent) and an intestinal deworming intervention (45 percent), but participation in these two programs is statistically similar across the three groups.

We also find that most schools lack basic school equipment and infrastructure. Ninety-three percent of schools use latrines instead of toilets, and only 56 percent have potable water. When we look at information on school buildings, our data shows that only 76 percent of school buildings are built exclusively for the school; the rest of the schools occupy either a church or a temporary building (8 percent), a hangar (7 percent), or other types of buildings (9 percent). The most common type of roof for buildings is sheet metal (76 percent) and most of the walls are made of concrete (88 percent). The differences observed between the three groups of schools in terms of school equipment and infrastructure are not statistically significant. Similarly, we did not find statistically significant differences between the three groups of schools in terms of the number of students per year of study or students per classroom.

Looking at management practices in Table C.9, we find that parents of students tend to participate in the management of the school, and most schools have a school management council (SMC). Around 90 percent of schools have a parents' committee (PC) and this committee met an average of two times between October 2014 and February 2015. Similarly, approximately 90 percent of schools have a school management council. About 70 percent of schools have teacher and student attendance control registries. We do not observe statistically significant differences in these variables across groups.

Throughout this paper, we use the Early Grade Reading Assessment (EGRA) to measure students' learning. EGRA is an orally administered assessment of foundational reading skills. This assessment aims to measure the most important components of good reading acquisition in terms of ability, speed, and fluency, in the early years of basic education through nine different subtasks. Each subtask is intended to measure a different competency needed to be able to read.[8] Throughout this paper, results from the EGRA will be our primary outcome of interest.[9] For simplicity, following Fazzio et al. (2020), we create a composite score based on the arithmetic mean of the child's scores in the EGRA subtasks. Within each section or subtask, all items were also given the same weight. We use this composite score as the main outcome of interest in the following sections. We also report individual test scores, scores per subtask, and the share of zero scores, in line with other work using EGRA (Eble et al., 2021; Dubeck and Gove, 2015). For individual subtasks, we focus on the percentage of correct answers. We tested the EGRA instruments for reliability before and after the first round of data collection. Our results confirm that the test is indeed reliable, with a Cronbach's alpha coefficient of 0.84.

Table 3.2 reports baseline results in the EGRA test for first-grade students in treated and control schools. Overall, there are no significant differences across groups in the subtasks measured by the EGRA nor on the composite score. The one exception is "listening comprehension," where we found that students in the control group performed better than those in the two groups that

[8]The nine competencies evaluated are: i) knowledge or recognition of letter names, ii) phonemic knowledge or awareness, iii) knowledge or recognition of letter sounds, iv) reading of familiar words (measured as number of words read in 60 seconds), v) reading of invented words (measured as number of words read in 60 seconds), vi) reading of a short story, vii) reading comprehension (based on short story), viii) listening comprehension, and ix) word dictation. A description of each subtask is included in Appendix Table C.1.

[9]The EGRA instruments used in this study were administered in Haitian Creole. These tools were developed by the World Bank in 2009 as part of a large effort to assess the reading performance of children in the departments of Nippes and the Artibonite (EdData II, 2014).

received the intervention. This gap may be explained by the fact that the Full and Light schools were ordered by the Ministry of National Education and Vocational Training (MENFP) to wait for the program before starting to teach oral reading. While these schools had to wait, the schools in the control group did not have this constraint and taught according to their usual timeline. Implementation delays in beneficiary schools may account for these rare deviations between the control and treatment group. We did not find statistically significant differences in the distribution of the composite score across groups, except when looking at differences between the Light and Control groups (see Appendix Figure C.1). This difference seems to favor the latter (this group's distribution is more skewed to the right).

According to the MENFP, all students must correctly read at least 30 Creole words per minute by the end of the 2nd grade. However, in our sample, less than 4 percent of students in the control group were able to read at least 30 words per minute of a Creole reading text at baseline. Overall, our data shows extremely low levels of pre-literacy skills among students halfway through the 1st grade, with a composite score of 12.8 out of 100 in the Creole reading assessments (see Table 3.3). The scores achieved at the subtask level are also cause for concern. Students can correctly read only 10 percent of the names of letters of a 100-letter table in 60 seconds ("letter recognition"), and can identify by sound 7 percent of letters of a 100-letter table in 60 seconds ("sound recognition"). When reading a text of 55 words ("oral reading fluency"), students can read correctly 5 words in 60 seconds (between 8 and 9 percent of their expected fluency), whereas the international standard for "good fluency" is 45 words per minute. This might explain why after reading the same 55-word text, the students only understand about 20 percent of the questions asked. The students also demonstrated difficulties in writing dictated words and letters: on a set of 5 letters and 3 words, they only managed to correctly write two letters and/or words (or 21-24 percent of the exercise). We do not find statistically significant differences among groups across all fluency measures.

We also estimate the share of students with zero scores in each subtask (see Table C.7 in the

Appendix). We find that in the majority of subtasks, most students have a zero score. For example, in the phonemic awareness task, between 70 and 78 percent of students in our sample are unable to identify at least one out of 45 words by their sound.

3.5 Implementation

The deployment of the M'ap Li Net Ale program can be separated into two main parts: a preparation phase and an implementation phase. During the preparation phase, the Commission Episcopale pour l'Education Catholique (CEEC), which was hired to support the MENFP in the implementation of the program, recruited key personnel for the integration of the methodology in schools, including 16 pedagogical trainers ("coaches"), 5 supervisors, and 9 community mobilizers. This stage also involved assessing the geographic and infrastructural conditions of the schools. Detailed visits were made to 290 participating schools, identifying accessibility challenges and grouping them into clusters for easier supervision. Each cluster was intended to include ten schools, although some clusters had to be reorganized when it was found that the distances between schools in certain regions were too vast for effective coordination.[10]

The preparation phase took place between December 2014 and January 2015. Training was a central component during this phase. After completion of staff training activities, a five-day training session was held for school principals and teachers in participating schools. Out of a total of 601 principals and teachers who were expected for training, 96 percent (577) were present.[11] During these sessions, all participants were trained in the use of the M'ap Li Net Ale method, which included phonemic awareness, reading comprehension strategies, and lesson planning. As part of the preparation phase, a total of 1,737 teacher guides and 9,911 student workbooks were distributed across the participating schools. In addition, CEEC organized a session for those who missed the initial training, with 18 participants completing a make-up session to ensure training

[10] Source: Programme "M ap li nèt ale": Rapport Final. CEEC.

[11] Online Appendix Figure C.3 shows the number of participants by region and type of school (Full or Light).

completion across schools.

During the implementation phase, the Full schools, which received the most comprehensive intervention, benefited from three remedial training sessions. The latter aimed to address identified weaknesses in teachers' work in different areas, including phonemic awareness, vocabulary and writing activities, management of the class, and other activities. After identifying important gaps in the application of the methodology during the first month of implementation (see Online Appendix Figure C.6), all principals and teachers in Full schools were invited to a remediation session, which took place over three days in April 2015. This session involved 260 participants, or 83 percent of the expected attendees. Participation rates were high across geographical departments, as shown in Online Appendix Figure C.4, ranging from 75 to 100 percent.

An important difference between Light and Full schools was the frequency of visits by external staff. Initially, Full schools were to receive visits from a pedagogical trainer (or coach) every two weeks and a supervisor once per month, while Light schools only received monthly visits from a supervisor. Supervisors monitored the intervention's implementation and supported school staff, while coaches focused on providing pedagogical assistance to teachers.[12] Coaches observed classroom instruction, gave feedback on how to improve their techniques, and provided guidance on how to apply the program more effectively. After each visit, they shared their notes with principals to ensure appropriate follow-up with teachers. In Light schools, the absence of biweekly visits from coaches meant supervisors and principals played a larger role. Principals, in particular, took over the role of coaching in Light schools.

After observing significant gaps in the application of the methodology during the first month of implementation, changes were made to the number of visits from coaches and supervisors in each type of school. As seen in Online Appendix Table C.2, the number of visits from coaching staff was increased to four per month in Full schools, and to one per month in a subset of Light

[12]Supervisors can also be thought of aa more senior figures who oversaw the relationships between coaches and teachers.

schools that showed large implementation difficulties. Given the large deficiencies identified among teachers, a decision was made to increase the length of coaching sessions during school visits. An implementation report found that during the beginning of the program, supervisors focused more on supporting principals and teachers in Light schools, which did not have the regular support of coaches, at the potential cost of neglecting supervision activities in Full schools. An implementation report states that this issue was quickly resolved after a meeting between the monitoring team and the supervisors.

Despite significant progress in teachers' understanding of the methodology, many schools still faced significant delays in the implementation of the program by the end of the first school year. The logistical challenges associated with carrying out the program were significant. Some schools, such as those in remote areas or areas affected by political unrest, required alternative delivery methods for materials, and were often affected by delays or cancellations of supervisor and coach visits.[13] Delays in teachers' payments were also reported as an important problem that affected teachers' motivation to continue with the application of the program. The implementation of extracurricular activities was similarly delayed due to payment issues.

Except for the cancellation of visits due to violence, which seemed to have affected a higher proportion of schools participating in the Light version, there is no evidence that these obstacles affected the two versions of the program differently. Moreover, anecdotal evidence suggests that a number of schools in the control group made and used copies of the M'ap Li Net Ale textbooks.[14] Considering this, our results should be interpreted as a lower-bound estimate of the true causal impact of the program, when compared to the control group. However, differences in results between the Light and Full program should be interpreted with caution given deviations from implementation protocols and mid-evaluation changes in program design.

[13] Visits to 9 school in the Light group were canceled due to heavy gunfire in the area. In these schools, directors were asked to put more emphasis on monitoring and following up on teachers' work in the classroom.

[14] Unfortunately, we do not have the exact number of schools in the control group that engaged in this practice.

3.5.1 Implementation Differences

By midline, participation levels were similar across Full and Light groups, at about 94-95 percent of all schools in each group (see Table C.13 in the Appendix). Six months after the intervention started, however, we already see important differences in how schools adopted and implemented each component of the intervention. A higher share of school staff in the Light group reported having received training in the M'ap Li Net Ale methodology by midline: about 86 percent of pedagogical principals in Light schools, compared to 65 percent in Full schools; 87 percent of administrative principals, compared to 74 in Full schools; and about 98 percent of teachers in Light schools compared to 92 percent in Full schools. The Light group also reported higher levels of use of the M'ap Li Net Ale methodology by teachers (98 percent compared to 91 percent in the Full group), and a higher likelihood of receiving the material to implement the program (96 percent of schools in the Light group compared to 83 percent in the Full group). Moreover, our data suggest that a slightly higher share of 1st graders in the Light group were benefiting from the program by midline (96 versus 93 percent in the Full group). These differences may be explained by the fact that Light schools did not have to wait for external coaches to start the intervention, and the training period was less lengthy than for Full schools. As mentioned, it is also possible that during the first few months supervisors may have compensated for the lack of coaches in Light schools.

By endline, most of these differences had disappeared, suggesting that schools in the Full group eventually caught up with Light schools in the components where implementation had been lagging at midline. For example, by endline a similar share of 1st grade teachers were trained in Light and Full schools (98-99 percent), as well as a similar number of staff per school (see Table C.14). Likewise, similar shares of principals in both types of schools reported that their first-year teachers had applied the M'ap Li Net Ale methodology to teach Creole (94-97 percent), and similar proportions of schools across groups reported having received the pedagogical materials for the program (92-93 percent).

3.6 Methodology

We estimate the impact of the program through simple Ordinary Least Squares (OLS) using the following equation:

$$\text{Score}_{gs} = \beta_1 + \beta_2\text{Light}_g + \beta_3\text{Full}_g + \epsilon_{gs}$$

In the equation provided, Score_{gs} represents the score of student s in school g. The variable Light_g serves as a dummy variable that is set to 1 if school g participated in the Light program and 0 otherwise. Similarly, Full$g$ is a dummy variable that takes the value of 1 if school g was treated by the Full program, and 0 otherwise. Lastly, ϵ_{gs} denotes the error term, which varies by student s and school g.

The randomization of schools into the Light, Full, and control groups was conducted through simple random assignment. We run this regression clustering standard errors at the school level, the unit of randomization. β_2 and β_3 estimate the average causal impact of each intervention. β_2 estimates the impact of the Light version, and β_3 corresponds to the impact of the Full version. We expect both coefficients to be statistically significant and positive.

3.7 Results

In this section, we present the main results of our analysis. We look at intermediate and final impacts of the Light and Full interventions on student learning outcomes, using data from midline and endline surveys.

Intermediate results presented in Table 3.4 suggest that by midline (early in students' 2nd grade school year) the intervention had a significant although marginal impact for both versions of the program. We found positive results in two out of nine subtasks assessed. Treated students in both Light and Full schools improved their scores in the phonemic awareness subtask by about 9.5 and 11.7 percentage points, respectively, relative to the control group. For the sound recognition subtask, treated students in the Light and Full programs increased their average score by about 1.9 and 1.7 percentage points relative to students in the control group (see Table 3.4). We do not observe significant impacts when we look at the composite score, our main outcome of interest (Table 3.4), or at fluency scores, at midline (Table 3.6).

By midline, we start noticing significant reductions in the share of students with zero scores (see Table 3.5). In phonemic awareness subtasks, the share of students with zero scores decreased by 12.9 percentage points (19.3 percent) in the Light group and by 14.1 percentage points (21.1 percent) in the Full group, relative to the control group. The share of students with zero scores on listening comprehension subtasks decreased by 1 percentage point (38 percent) in the Light group and 1.2 percentage points (47 percent) in the Full group, relative to the control group.

Looking at endline results in Table 3.7, we find that the program had large and positive effects on most reading measures. Panel A shows results for the second cohort (1st grade), which benefited from the intervention in the 2015-16 school year. The first cohort (2nd grade), shown in Panel B, benefited from the intervention for two consecutive school years. Interestingly, effects across the two types of treatment (Light and Full) look very similar across all subtasks. Overall, our estimates suggest that the programs were equally effective in improving students' learning outcomes. Effects for the cohort that benefited from the program during two consecutive years are two to three times as large as those of the cohort that benefited for only one year. These effects mirror those of D'Agostino et al. (2024), who find larger effects for a 2nd grade cohort who received 2 years of treatment of a close version of this intervention. Moreover, as in D'Agostino et al. (2024), only students who received the intervention for two consecutive school years show gains in reading comprehension.

Table 3.7 shows that treated students in the first cohort in the Light and Full programs scored 7.2 (26.8 percent) and 6.8 percentage points (25.4 percent) higher than the control group in the composite score. Treated students in the second cohort scored 2.4 (16 percent) and 1.9 percentage points (13 percent) higher than the control group in this measure. We did not find statistically significant differences in this score between the Light and Full schools. However, this impact is not homogeneous across the different topics evaluated in the EGRA test. The largest impacts were found on phonemic awareness and sound recognition subtasks for the first cohort, which increased by 32-34 percentage points (136-144 percent) and by 8.7-8.9 percentage points (62.8-64.3 percent), relative to the control group. For the second cohort, impacts also seem largely driven by improvements in these two dimensions. We did not find statistically significant differences between these two groups in any of the subtasks, which suggests that the Light program achieved the same level of effectiveness as the Full program.

In Table 3.8, we present fluency scores at endline for the first cohort. Fluency measures for the sound recognition subtask improved by 8.9 percentage points (64.2 percent) in the Light program and 8.6 percentage points (62 percent) in the Full program. Similarly, fluency in reading familiar words improved by 3 percentage points (23.7 percent) in the Light program and 2.5 percentage points (19.6 percent) in the Full program. We did not find, however, any improvements in our measure of fluency in oral reading or letter recognition.

We only observe significant improvements in the share of students with zero scores for one of the nine subtasks measured at endline (Table 3.9). With respect to the control group, students in the Light group were 33.5 percentage points (52.5 percent) and 36.3 percentage points (56.9 percent) less likely to have a score of zero in the phonemic awareness subtask. We did not find statistically significant improvements in other subtasks. These high shares of zero scores suggest that improvements in reading skills were not homogeneous across students, and have also been found for other similarly comprehensive reading interventions implemented in low-income settings, including in Ghana and Haiti (Guzmán et al., 2021; D'Agostino et al., 2024; USAID, 2019).

We present results in terms of standard deviations in Appendix Tables C.11 and C.12 at midline and endline. The results are in line with the estimates previously presented. At endline, we find effects of 0.33-0.34 standard deviations on our composite score.

3.8 Cost-Effectiveness Analysis

In this section, we estimate the impact of the program per USD 100 spent and compare it with other interventions aiming to improve learning outcomes in developing countries.

For the sake of this exercise, we use the program's impact in terms of standard deviations (SD) to measure and compare its cost-effectiveness, as this is the standard measure used in the literature. After two years, the Light intervention cost approximately 253 dollars per student (approximately 160 dollars for the first year, and 93 dollars for the second).

The Full version is estimated at 298 dollars (198 dollars for the first year and 100 dollars for the second year). Considering potential savings from scaling up the intervention, we consider the amount of 298 dollars an upper bound of the per-student costs of implementing the Full program at a larger scale. Differences in costs between the Light and Full program can be largely attributed to the salaries of external coaches.

For both interventions, we find effects of approximately 0.33-0.34 SD over two years in our composite reading score, our main outcome of interest.[15] This is equivalent to 0.11 SD per 100 dollars spent in the case of the Full version, and 0.13 SD per 100 dollars spent for the Light version, assuming a linear dose-response relationship. To put this in context, this is almost four times as effective as an intervention giving conditional cash transfers in Malawi (Baird et al., 2011), or twice as large as the effect of a Computer Assisted Learning (CAL) after-school program in India (Muralidharan et al., 2019).

With per-student costs ranging from 253 to 298 US dollars, this intervention is not among the

[15] In the education literature, an increase of more than 0.2 SD is considered a large effect (Kraft, 2020).

most cost-effective to improve student outcomes in the economic literature.[16] Nonetheless, non-linearities in the program's impact, as evidenced when comparing midline and endline results, suggest that it could yield greater long-term improvements in reading skills.

Comparing the Full and Light versions of the program, we see that the latter achieves virtually identical results at a lower cost. Or put differently, achieves larger results per 100 dollars spent (0.13 SD vs 0.11 SD per 100 dollars spent). In this sense, the Light version seems considerably more cost-effective than the Full version. However, the cost of the program is relatively high when considering the Haitian government's annual expenditure on education. For reference, in 2015 the government gave USD 90 per student to government-accredited schools to cover education expenses through the Tuition Waiver Program (World Bank, 2015).

3.9 Discussion

While our results suggest that the same level of effectiveness can be achieved at a lower cost when coaching activities are led by school principals rather than an external team, these results could be compromised if other factors affecting results, beyond the type of supervision, differed across the Full and Light groups. We examined data around implementation from midline and endline surveys to assess the extent to which the different components of the intervention were differentially implemented across groups.

The results suggest that implementation differences, particularly during the first months of the program, may have given Light group schools a slight edge over the Full group. Endline data, however, makes evident that these early differences eventually disappeared. By the end of the program, both groups had similarly complied with the different components of the intervention, except for the type of coaching, which was designed to be the only difference between the two

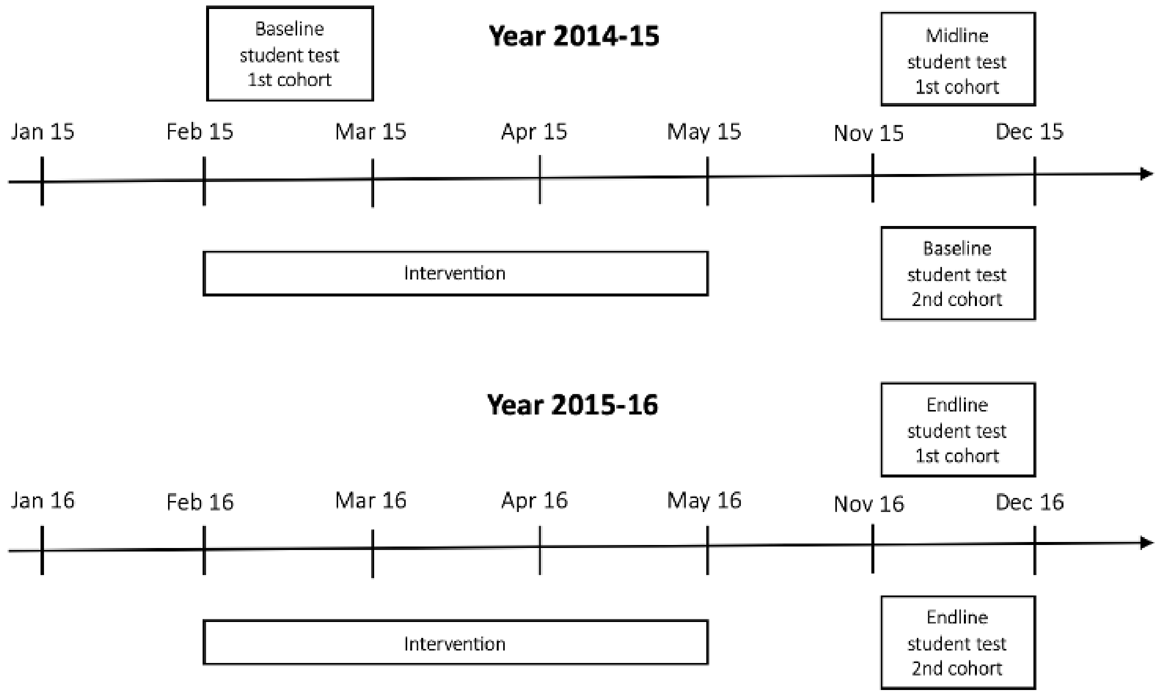
[16]For example, interventions relying on the use of technology have proven to be substantially cheaper and highly effective, at USD 75 in India (Muralidharan et al., 2019), or USD 43-56 in El Salvador (Büchel et al., 2022), with effect sizes ranging from 0.21 to 0.37 SD. However, the context in which this study takes place lacks the appropriate infrastructure required for this type of intervention.

groups. Although this does not constitute enough evidence to confidently conclude that early program differences did not contribute to observed results, the lack of significant differences across group outcomes at midline supports the hypothesis that these early differences were not substantial to each intervention's effectiveness.

In sum, the Light version was able to achieve the same effects without relying on an external team for coaching. Three possible mechanisms behind these results are flexibility, ownership, and accountability within schools. Effective teachers must adapt and alter their instruction depending on their students' changing needs, and teachers in the Light version, where coaching efforts were led by their principals, who they had easier access to, had more flexibility to do so. It is also possible that external coaching may have negatively affected teachers' sense of ownership and accountability around the program, relative to internal (principal-led) coaching. Indeed, we find that in the following school year, 1st grade teachers that had been trained in the program the previous year were much more likely to continue applying the methodology in schools in the Light group than in schools in the Full group (97 versus 67 percent), although this difference decreases substantially by endline. The new organizational relationships put in place by the Light program could have increased teachers' motivation to permanently adopt the program's methodology. By giving direct responsibilities to principals, the program was also more likely to change the day-to-day practices of principals, leading to more sustained effects over the long run. These are important areas for future research.

These results contribute to a growing body of literature aiming to identify effective and sustainable approaches for implementing large-scale early-grade reading programs that incorporate key elements like coaching. One approach focuses on leveraging scarce high-skilled human resources through virtual coaching. While intuitively appealing, results from South Africa suggest that this approach may not significantly reduce costs and may not have as lasting an impact (Cilliers et al., 2021). The Light version of the program evaluated in this paper takes a different approach, attempting to use the human resources already available in schools to provide coaching. While we do not have data to assess the long-term impacts of the program, the short-term results presented above suggest that this may be a promising approach even in contexts as challenging as Haiti.

Figure 3.1: Timeline for Data Collection and Program Implementation



Notes: The first cohort is composed of students enrolled in the 1st grade in the 2014-15 school year. The second cohort includes students enrolled in the 1st grade in the 2015-16 school year.

Table 3.1: Schools and Students by Geographic Department at Baseline

	Schools				Students			
	Control	Full	Light	Total	Control	Full	Light	Total
Department	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Grande-Anse	30	25	28	83	489	431	445	1365
Nippes	61	67	63	191	848	989	913	2750
Ouest	5	14	15	34	89	235	275	599
Nord-Ouest	24	26	27	77	429	473	436	1338
Total	120	132	133	385	1855	2128	2069	6052

Source: authors' own estimates. This sample covers students who were enrolled in the first grade in the 2014-2015 school year.

Table 3.2: Student Characteristics at Baseline

	Control (1)	Light (2)	Full (3)	Difference L-C (4)	Difference F-C (5)	Difference F-L (6)
Age	7.33	7.19	7.23	-0.14	-0.10	0.04
Gender-boy	0.55	0.53	0.53	-0.02	-0.02	0.00
Pupil lives with his parents						
Dad	0.76	0.74	0.73	-0.02	-0.03	-0.01
Mom	0.84	0.85	0.85	0.01	0.01	0.00
Both	0.68	0.68	0.67	0.00	-0.01	-0.01
Student attended preschool	0.72	0.76	0.80	0.04***	0.08*	0.04
Student ate breakfast before going to school	0.73	0.71	0.68	-0.02**	-0.05	-0.03
Student's school equipment						
Pencil	0.78	0.72	0.70	-0.06***	-0.08	-0.02***
Pen	0.04	0.05	0.04	0.01	0.00	-0.01
Chalk	0.07	0.04	0.05	-0.03	-0.02	0.01*
Slate	0.07	0.05	0.06	-0.02	-0.01	0.01
Notebook	0.79	0.73	0.75	-0.06*	-0.04	0.02**
Family property and utilities						
Electricity	0.33	0.33	0.32	0.00	-0.01	-0.01
Radio	0.50	0.48	0.48	-0.02	-0.02	0.00
Television	0.20	0.20	0.20	0.00	0.00	0.00
Fridge	0.11	0.10	0.11	-0.01	0.00	0.01
Computer	0.06	0.04	0.06	-0.02	0.00	0.02**
Bicycle	0.11	0.11	0.12	0.00	0.01	0.01
Motorbike	0.12	0.11	0.13	-0.01	0.01	0.02
Car	0.06	0.06	0.06	0.00	0.00	0.00

Source: Baseline implementation report.

Table 3.3: EGRA Results at Baseline as Percentage of Correct Answers

Task	Control (1)	Light (2)	Full (3)	Difference L-C (4)	Difference F-C (5)	Difference F-L (6)
Letter recognition	10.13	10.05	9.57	-0.08	-0.56	-0.48
Phonemic awareness	14.14	16.67	18.31	2.52	4.17*	1.65
Sound recognition	6.83	7.45	6.62	0.61	-0.22	-0.83
Familiar word reading	7.27	8.09	6.83	0.81	-0.44	-1.25
Reading invented words	5.94	7.11	5.67	1.17	-0.27	-1.43
Oral reading fluency	9.17	9.82	8.32	0.65	-0.85	-1.50
Reading comprehension	5.14	4.89	5.64	-0.25	0.50	0.75
Listening comprehension	34.26	29.74	33.32	-4.52***	-0.95	3.57***
Word dictation	23.54	22.30	21.36	-1.24	-2.17	-0.94
Overall composite score	12.94	12.90	12.85	-0.04	-0.09	-0.05

Notes: *p<0.10, **p<0.05, ***p<0.01. Standard errors are clustered at the school level. Each category has been scaled to reflect the percentage of correct answers, except for the following categories: Familiar words reading, Invented words reading, and Oral reading fluency. In these cases, the scores have been scaled so that 100 represents students who were able to read at least 55 words within 60 seconds. N(control)=1855, N(Light)=2069, N(Full)=2128.

Table 3.4: EGRA Results at Midline as Percentage of Correct Answers

Task	Control (1)	Light (2)	Full (3)	Difference L-C (4)	Difference F-C (5)	Difference F-L (6)
Letter recognition	19.60	18.77	19.11	-0.82	-0.49	0.33
Phonemic awareness	19.11	28.63	30.84	9.53***	11.73***	2.20
Sound recognition	9.95	11.81	11.67	1.85**	1.72**	-0.13
Familiar word reading	15.39	16.44	16.32	1.05	0.93	-0.12
Reading invented words	10.02	10.88	10.80	0.85	0.78	-0.07
Oral reading fluency	21.11	21.95	21.28	0.85	0.17	-0.68
Reading comprehension	8.65	9.50	8.90	0.86	0.26	-0.60
Listening comprehension	63.95	63.05	65.84	-0.90	1.89*	2.79**
Dictated words	41.82	40.87	40.51	-0.95	-1.31	-0.36
Overall composite score	23.29	24.66	25.03	1.37	1.74	0.37

Notes: *p<0.10, **p<0.05, ***p<0.01. Standard errors are clustered at the school level. This sample only covers the first cohort of students (students enrolled in the 2nd grade in the 2015-16 school year). N(control)=1855, N(Light)=2069, N(Full)=2128.

Table 3.5: Additional EGRA Results at Midline: Percent with Zero Scores

Task	Control (1)	Light (2)	Full (3)	Difference L-C (4)	Difference F-C (5)	Difference F-L (6)
Letter recognition	9.89	11.81	11.13	1.92	1.23	-0.68
Phonemic awareness	66.91	53.99	52.77	-12.92***	-14.14***	-1.22
Sound recognition	18.68	19.06	16.98	0.38	-1.70	-2.08
Familiar word reading	42.65	43.82	42.78	1.17	0.13	-1.04
Reading invented words	67.82	64.80	66.98	-3.02	-0.83	2.19
Oral reading fluency	39.34	40.36	40.37	1.02	1.04	0.02
Reading comprehension	76.95	75.88	77.02	-1.07	0.08	1.14
Listening comprehension	2.55	1.60	1.36	-0.95**	-1.18***	-0.23
Word dictation	12.01	13.86	15.08	1.85	3.07*	1.21
Overall composite score	0.29	0.41	0.27	0.12	-0.02	-0.14

Notes: *p<0.10, **p<0.05, ***p<0.01. Standard errors are clustered at the school level. Estimates for each task represent the share of students with a zero score. N(control)=2046, N(Light)=2064, N(Full)=2039. This sample only covers the first cohort of students (students enrolled in the 2nd grade in the 2015-16 school year).

Table 3.6: Additional EGRA Results at Midline: Fluency Scores

Task	Control (1)	Light (2)	Full (3)	Difference L-C (4)	Difference F-C (5)	Difference F-L (6)
Letter recognition	20.14	18.84	19.15	-1.30	-1.00	0.31
Sound recognition	10.40	11.87	11.73	1.47	1.34	-0.13
Familiar word reading	8.10	8.70	8.62	0.60	0.53	-0.07
Reading invented words	5.16	5.58	5.56	0.42	0.40	-0.02
Oral reading fluency	12.60	13.27	12.84	0.67	0.24	-0.44

Notes: *p<0.10, **p<0.05, ***p<0.01. Standard errors are clustered at the school level. Fluency scores are respectively computed as: i) total number of letters for which the student correctly identifies the name in one minute; ii) number of letters sounds correctly identified in one minute; iii) number of familiar words correctly read in one minute; iv) number of coined words correctly read in one minute; and v) number of words correctly read in one minute from a short text. Due to lack of response time data, fluency scores are not measured for the following subtasks: phonemic awareness, oral reading fluency, reading comprehension, listening comprehension, and word dictation. N(control)=2046, N(Light)=2064, N(Full)=2039. This sample only covers the first cohort of students (students enrolled in the 2nd grade in the 2015-16 school year).

Table 3.7: EGRA Results at Endline as Percentage of Correct Answers

Task	Control (1)	Light (2)	Full (3)	Difference L-C (4)	Difference F-C (5)	Difference F-L (6)
Panel A. 1st Grade (2nd Cohort)						
Letter recognition	14.09	13.19	12.49	-0.91	-1.60	-0.69
Phonemic awareness	15.41	32.30	33.99	16.89***	18.58***	1.68
Sound recognition	9.23	12.11	11.14	2.88***	1.91**	-0.96
Familiar word reading	9.72	10.67	9.52	0.95	-0.20	-1.14
Reading invented words	7.44	7.56	6.67	0.11	-0.78	-0.89
Oral reading fluency	8.50	9.07	8.10	0.57	-0.40	-0.97
Reading comprehension	4.25	4.72	4.41	0.47	0.16	-0.31
Listening comprehension	38.17	40.22	40.76	2.05	2.59*	0.54
Word dictation	29.00	27.75	25.99	-1.25	-3.01	-1.76
Overall composite score	15.09	17.51	17.01	2.42**	1.92*	-0.50
Panel B. 2nd Grade (1st Cohort)						
Letter recognition	26.20	26.32	25.43	0.12	-0.78	-0.89
Phonemic awareness	23.48	55.51	57.48	32.03***	34.00***	1.97
Sound recognition	13.78	22.67	22.45	8.88***	8.67***	-0.21
Familiar word reading	23.37	28.59	27.79	5.21**	4.42*	-0.79
Reading invented words	17.36	20.62	19.59	3.26	2.23	-1.04
Oral reading fluency	22.33	27.63	26.44	5.29**	4.10	-1.19
Reading comprehension	13.41	17.15	16.33	3.74**	2.92*	-0.82
Listening comprehension	52.22	54.16	55.65	1.94	3.43**	1.49
Word dictation	49.66	53.97	52.22	4.31**	2.56	-1.74
Overall composite score	26.87	34.07	33.71	7.20***	6.84***	-0.36

Notes: *p<0.10, **p<0.05, ***p<0.01. Standard errors are clustered at the school level. Panel A shows the second cohort of students: students who were enrolled in the first grade in 2015-16. Treated students in this cohort received the intervention only during the 2015-16 school year. Panel B shows the first cohort: students who were enrolled in the second grade in 2015-16 and who benefited from the intervention for two consecutive school years (2014-15 and 2015-16). Sample size for the first grade: N(control)=2001, N(Light)=2061, N(Full)=2043; for the second grade: N(control)=2062, N(Light)=2178, N(Full)=2175.

Table 3.8: Additional EGRA Results at Endline: Fluency Scores

Task	Control (1)	Light (2)	Full (3)	Difference L-C (4)	Difference F-C (5)	Difference F-L (6)
Letter recognition	26.29	26.36	25.45	0.07	-0.84	-0.92
Sound recognition	13.91	22.83	22.53	8.92***	8.62***	-0.30
Familiar word reading	12.60	15.58	15.07	2.98**	2.47*	-0.51
Reading invented words	9.05	10.73	10.14	1.68	1.09	-0.59
Oral reading fluency	10.69	12.41	11.48	1.72	0.80	-0.92

Notes: *p<0.10, **p<0.05, ***p<0.01. Standard errors are clustered at the school level. Fluency scores are respectively computed as: (i) total number of letters out of 100 for which the student correctly identifies the name in one minute; (ii) number of letters sounds correctly identified in one minute; (iii) number of familiar words correctly read in one minute; (iv) number of coined words correctly read in one minute; and (v) number of words correctly read in one minute from a short text. These scores are only available for the first cohort (students enrolled in the second grade in the 2015-16 school year). N(control) = 2062, N(Light) = 2178, N(Full) = 2175.

Table 3.9: Additional EGRA Results at Endline: Percent with Zero Scores

Task	Control (C) (1)	Light (L) (2)	Full (F) (3)	Difference L-C (4)	Difference F-C (5)	Difference F-L (6)
Letter recognition	10.52	10.10	11.22	-0.42	0.69	1.12
Phonemic awareness	63.82	30.35	27.54	-33.47***	-36.28***	-2.81
Sound recognition	11.45	9.87	8.78	-1.57	-2.66*	-1.09
Familiar word reading	37.20	37.65	36.74	0.45	-0.46	-0.91
Reading invented words	43.40	45.22	46.62	1.82	3.22	1.40
Oral reading fluency	51.79	48.85	49.20	-2.94	-2.60	0.34
Reading comprehension	69.88	63.82	65.01	-6.06*	-4.87	1.19
Listening comprehension	9.99	9.46	8.41	-0.53	-1.58	-1.04
Word dictation	8.20	8.03	9.56	-0.16	1.37	1.53
Overall composite score	0.82	0.78	0.60	-0.04	-0.23	-0.18

Notes: *p¡0.10, **p¡0.05, ***p¡0.01. Standard errors are clustered at the school level. Estimates for each task represent the share of students with a zero score. N(control) = 2062, N(Light) = 2178, N(Full) = 2175. These scores are only available for the first cohort of students (students enrolled in the second grade in the 2015-16 school year).

Conclusion

Improving student learning remains a core objective of education policy, yet the most effective strategies to achieve this goal often depend on the specific context. This dissertation has examined three distinct interventions implemented in Haiti, the Dominican Republic, and the United States, each offering insights into how educational outcomes can be improved through different approaches.

While the interventions vary in design and setting—from structured pedagogy in a fragile system to adaptive technology in middle-income classrooms and school choice reforms in a high-income context—they share a common focus on understanding how changes in key inputs—such as instructional methods, school organization, and teacher support—affect student outcomes within the educational production function. The findings demonstrate that there is no universal blueprint for success; rather, effective interventions must be tailored to local conditions and capacities. Importantly, the chapters show that both short-term academic outcomes and longer-term indicators such as college enrollment and completion can be influenced through well-designed approaches.

Taken together, the evidence presented in this dissertation highlights the importance of context-sensitive, evidence-based approaches to improving education systems. It contributes to a deeper understanding of what works, for whom, and under what circumstances, contributing to ongoing efforts to advance educational equity and effectiveness.

References

Abadie, A. (2002). Bootstrap tests for distributional treatment effects in instrumental variables models. *Journal of the American Statistical Association 97*(457), 284–292.

Abdulkadiroğlu, A., J. Angrist, S. Dynarski, T. J. Kane, and P. Pathak (2011). Accountability and flexibility in public schools: Evidence from Boston's charters and pilots. *The Quarterly Journal of Economics 126*(2), 699–748.

Adelman, M., J. D. Baron, M. Blimpo, D. K. Evans, A. Simbou, and N. Yarrow (2015). Why do students learn so little? seeking answers inside haiti's classrooms. Technical report, World Bank. Accessed: 2023-04-23.

Alan, S., T. Boneva, and S. Ertac (2019). Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. *The Quarterly Journal of Economics 134*(3), 1121–1162.

Angel-Urdinola, D., C. Avitabile, and M. Chinen (2023). Can digital personalized learning for mathematics remediation level the playing field in higher education? experimental evidence from ecuador.

Angrist, J. and V. Lavy (2002). New evidence on classroom computers and pupil learning. *The Economic Journal 112*(482), 735–765.

Angrist, J., P. Pathak, and C. Walters (2013). Explaining charter school effectiveness. *American Economic Journal: Applied Economics 5*(4), 1–27.

Angrist, J. D., S. R. Cohodes, S. M. Dynarski, P. A. Pathak, and C. R. Walters (2016). Stand

and deliver: Effects of Boston's charter high schools on college preparation, entry, and choice. *Journal of Labor Economics 34*(2), 275–318.

Angrist, N., P. Bergman, C. Brewster, and M. Matsheng (2020). Stemming learning loss during the pandemic: A rapid randomized trial of a low-tech intervention in botswana. *Available at SSRN 3663098*.

Antoninis, M., B. Alcott, S. Al Hadheri, D. April, B. Fouad Barakat, M. Barrios Rivera, Y. Baskakova, M. Barry, Y. Bekkouche, D. Caro Vasquez, et al. (2023). Global education monitoring report 2023: Technology in education: A tool on whose terms?

Araya, R., E. A. Ortiz, N. L. Bottan, and J. P. Cristia (2019). Does gamification in education work?: Experimental evidence from chile.

Athey, S. and G. W. Imbens (2017). The econometrics of randomized experiments. In *Handbook of economic field experiments*, Volume 1, pp. 73–140. Elsevier.

Augustine, C. H., J. S. McCombs, J. F. Pane, H. L. Schwartz, J. Schweig, A. McEachin, and K. Siler-Evans (2016). *Learning from Summer: Effects of Voluntary Summer Learning Programs on Low-Income Urban Youth. RAND Summer Learning Series. Research Report. RR-1557-WF.* ERIC.

Autor, D., A. Dube, and A. McGrew (2023). The unexpected compression: Competition at work in the low wage labor market. Technical report, National Bureau of Economic Research.

Baird, S., C. McIntosh, and B. Özler (2011). Cash or condition? Evidence from a cash transfer experiment. *The Quarterly Journal of Economics 126*(4), 1709–1753.

Banerjee, A., E. Duflo, N. Goldberg, D. Karlan, R. Osei, W. Parienté, J. Shapiro, B. Thuysbaert, and C. Udry (2015). A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science 348*(6236).

Baron, J., M. Adelman, and D. Evans (2016, June). Why school enrollment is not enough: A look inside haiti's classrooms. Education for Global Development, The World Bank Blog.

Barrow, L., L. Markman, and C. E. Rouse (2009). Technology's edge: The educational benefits of computer-aided instruction. *American Economic Journal: Economic Policy 1*(1), 52–74.

Bengali, L., M. Sander, R. G. Valletta, and C. Zhao (2023). Falling college wage premiums by race and ethnicity. *FRBSF Economic Letter 2023*(22), 1–6.

Black, S. E., J. T. Denning, and J. Rothstein (2023). Winners and losers? the effect of gaining and losing access to selective colleges on education and labor market outcomes. *American Economic Journal: Applied Economics 15*(1), 26–67.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher 13*(6), 4–16.

Borman, G. D., G. M. Hewes, L. T. Overman, and S. Brown (2002). Comprehensive school reform and student achievement: A meta-analysis.

Borman, G. D., R. E. Slavin, A. Cheung, A. M. Chamberlain, N. A. Madden, and B. Chambers (2005a). Success for all: First-year results from the national randomized field trial. *Educational Evaluation and Policy Analysis 27*(1), 1–22.

Borman, G. D., R. E. Slavin, A. C. Cheung, A. M. Chamberlain, N. A. Madden, and B. Chambers (2005b). The national randomized field trial of success for all: Second-year outcomes. *American Educational Research Journal 42*(4), 673–696.

Borman, G. D., R. E. Slavin, A. C. Cheung, A. M. Chamberlain, N. A. Madden, and B. Chambers (2007). Final reading outcomes of the national randomized field trial of success for all. *American Educational Research Journal 44*(3), 701–731.

Bruns, B. and J. Luque (2014). *Great teachers: How to raise student learning in Latin America and the Caribbean*. World Bank Publications.

Büchel, K., M. Jakob, C. Kühnhanss, D. Steffen, and A. Brunetti (2022). The relative effectiveness of teachers and learning software: Evidence from a field experiment in el salvador. *Journal of labor economics 40*(3), 737–777.

Carlana, M. and E. La Ferrara (2021). Apart but connected: Online tutoring and student outcomes during the covid-19 pandemic.

Chabrier, J., S. Cohodes, and P. Oreopoulos (2016). What can we learn from charter school lotteries? *Journal of Economic Perspectives 30*(3), 57–84.

Chetty, R., J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan (2011). How does your kindergarten classroom affect your earnings? evidence from project star. *The Quarterly journal of economics 126*(4), 1593–1660.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014a). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American economic review 104*(9), 2593–2632.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014b). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American economic review 104*(9), 2633–2679.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014c). Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood. *American economic review 104*(9), 2633–2679.

Chingos, M. M. and P. E. Peterson (2011). It's easier to pick a good teacher than to train one: Familiar and new results on the correlates of teacher effectiveness. *Economics of Education Review 30*(3), 449–465.

Cilliers, J., B. Fleisch, N. M. Kotzé, S. Taylor, and T. Thulare (2021). Can virtual replace in-person coaching? experimental evidence on teacher professional development and student learning. *Unpublished*.

Cohodes, S. and J. J. Feigenbaum (2021, September). Why does education increase voting? Evidence from Boston's charter schools. Working Paper 29308, National Bureau of Economic Research.

Cohodes, S. and S. Roy (2024). Thirty years of charter schools: What does lottery-based research tell us? *Journal of School Choice 0*(0), 1–42.

Cohodes, S. R. and J. S. Goodman (2014). Merit aid, college quality, and college completion: Massachusetts' adams scholarship as an in-kind subsidy. *American Economic Journal: Applied Economics 6*(4), 251–285.

Cohodes, S. R. and K. S. Parham (2021). Charter schools' effectiveness, mechanisms, and competetive influence. *National Bureau of Economic Research Working Paper 28477*.

Cohodes, S. R., E. M. Setren, and C. R. Walters (2021). Can successful schools replicate? Scaling up Boston's charter school sector. *American Economic Journal: Economic Policy 13*(1), 138–67.

Crawford, M., M. Barron, M. Gutierrez, and E. Ding (2020, September). Universalizing basic literacy: how to get every child reading?

Davis, M. and B. Heller (2019). No excuses charter schools and college enrollment: New evidence from a high school network in Chicago. *Education Finance and Policy 14*(3), 414–440.

De Hoyos, R., A. J. Ganimian, and P. A. Holland (2021). Teaching with the test: experimental evidence on diagnostic feedback and capacity building for public schools in argentina. *The World Bank Economic Review 35*(2), 499–520.

DeAngelo, L., R. Franke, S. Hurtado, J. H. Pryor, and S. Tran (2011). Completing college: Assessing graduation rates at four-year institutions.

DeGraff, M. (2016). Mother-tongue books in haiti: The power of kreyòl in learning to read and in reading to learn. *Prospects 46*(3), 435–464.

Demers, A., I. Nichols-Barrer, E. Steele, M. Bartlett, and P. Gleason (2017). Long-term impacts of KIPP middle and high schools on college enrollment, persistence, and attainment. Technical report, Mathematica.

Dobbie, W. and R. Fryer (2013). Getting beneath the veil of effective schools: Evidence from New York City. *American Economic Journal: Applied Economics 5*(4), 28–60.

Dobbie, W. and R. Fryer (2015). The medium-term impacts of high-achieving charter schools. *Journal of Political Economy 123*(5), 985–1037.

Dobbie, W. and R. G. Fryer (2020). Charter schools and labor market outcomes. *Journal of Labor Economics 38*(4), 915–957.

Doignon, J. P. and J. C. Falmagne (1985). Spaces for the assessment of knowledge. *International Journal of Man-Machine Studies 23*(2), 175–196.

Dowker, A., A. Sarkar, and C. Y. Looi (2016). Mathematics anxiety: What have we learned in 60 years? *Frontiers in psychology 7*, 508.

Dubeck, M. M. and A. Gove (2015). The early grade reading assessment (EGRA): Its theoretical foundation, purpose, and limitations. *International Journal of Educational Development 40*, 315–322.

Duckworth, A. L., C. Peterson, M. D. Matthews, and D. R. Kelly (2007). Grit: perseverance and passion for long-term goals. *Journal of personality and social psychology 92*(6), 1087.

Duncan, G. J., C. J. Dowsett, A. Claessens, K. Magnuson, A. C. Huston, P. Klebanov, L. S. Pagani, L. Feinstein, M. Engel, J. Brooks-Gunn, H. Sexton, K. Duckworth, and C. Japel (2007). School readiness and later achievement. *Developmental Psychology 43*(6), 1428–1446.

Dynarski, S., D. Hubbard, B. Jacob, and S. Robles (2018). Estimating the effects of a large for-profit charter school operator. Technical report, National Bureau of Economic Research.

Dynarski, S., A. Nurshatayeva, L. C. Page, and J. Scott-Clayton (2023). Addressing nonfinancial barriers to college access and success: Evidence and policy implications. Volume 6 of *Handbook of the Economics of Education*, pp. 319–403. Elsevier.

D'Agostino, T., D. B. Guzmán, P. Perrin, A. Liberiste-Osirus, and K. Schuenke-Lucien (2024). Explaining variation in treatment effects: An impact evaluation and mixed-methods study of variation in early grade reading program effects. *Comparative Education Review 68*(1), 85–112.

Eble, A., C. Frost, A. Camara, B. Bouy, M. Bah, M. Sivaraman, C. Jayanty, T. Brady, P. Gawron, S. Vansteelandt, and D. 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.

EdData II (2014). Policy brief: Early grade reading assessment (egra) in haiti. Technical report, RTI International. Accessed: 2024-10-11.

Epple, D., R. Romano, and R. Zimmer (2016). Charter schools: A survey of research on their characteristics and effectiveness. In *Handbook of the Economics of Education*, Volume 5, pp. 139–208. Elsevier.

Evans, D. K. and F. Yuan (2022). How big are effect sizes in international education studies? *Educational Evaluation and Policy Analysis 44*(3), 532–540.

Falmagne, J.-C., M. Koppen, M. Villano, J.-P. Doignon, and L. Johannesen (1990). Introduction to knowledge spaces: How to build, test, and search them. *Psychological Review 97*(2), 201–224.

Fazzio, I., A. Eble, R. L. Lumsdaine, P. Boone, B. Bouy, P. T. J. Hsieh, C. Jayanty, S. Johnson, and

A. F. Silva (2020). Large learning gains in pockets of extreme poverty: Experimental evidence from guinea bissau. Technical Report w27799, National Bureau of Economic Research.

Ferguson, R. F. and C. Danielson (2015). How framework for teaching and tripod 7cs evidence distinguish key components of effective teaching. *Designing teacher evaluation systems: New guidance from the measures of effective teaching project*, 98–143.

Frühe, B., A.-K. Allgaier, K. Pietsch, M. Baethmann, J. Peters, S. Kellnar, A. Heep, S. Burdach, D. von Schweinitz, and G. Schulte-Körne (2012). Children's depression screener (child-s): development and validation of a depression screening instrument for children in pediatric care. *Child Psychiatry & Human Development 43*, 137–151.

Ganimian, A. J., K. Muralidharan, and C. R. Walters (2021). Augmenting state capacity for child development: Experimental evidence from india. Technical report, National Bureau of Economic Research.

Ge, S., E. Isaac, and A. Miller (2022). Elite schools and opting in: Effects of college selectivity on career and family outcomes. *Journal of Labor Economics 40*(S1), S383–S427.

Gershenson, S. (2016). Linking teacher quality, student attendance, and student achievement. *Education Finance and Policy 11*(2), 125–149.

Gleason, P., M. Clark, C. C. Tuttle, and E. Dwoyer (2010, June). The evaluation of charter school impacts: Final report. NCEE 2010-4029, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, Washington, DC.

Glewwe, P. and K. Muralidharan (2016). Improving education outcomes in developing countries: Evidence, knowledge gaps, and policy implications. In *Handbook of the Economics of Education*, Volume 5, pp. 653–743. Elsevier.

Goodman, J., M. Hurwitz, and J. Smith (2017). Access to 4-year public colleges and degree completion. *Journal of Labor Economics 35*(3), 829–867.

Green, D. P., P. M. Aronow, D. E. Bergan, P. Greene, C. Paris, and B. I. Weinberger (2011). Does knowledge of constitutional principles increase support for civil liberties? Results from a randomized field experiment. *The Journal of Politics 73*(2), 463–476.

Guryan, J., J. Ludwig, M. P. Bhatt, P. J. Cook, J. M. Davis, K. Dodge, G. Farkas, R. G. Fryer Jr, S. Mayer, H. Pollack, et al. (2023). Not too late: Improving academic outcomes among adolescents. *American Economic Review 113*(3), 738–765.

Guzmán, J. C., K. Schuenke-Lucien, A. J. D'Agostino, M. Berends, and A. J. Elliot (2021). Improving reading instruction and students' reading skills in the early grades: Evidence from a randomized evaluation in haiti. *Reading Research Quarterly 56*(1), 173–193.

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education review 30*(3), 466–479.

Hanushek, E. A., G. Schwerdt, S. Wiederhold, and L. Woessmann (2015). Returns to skills around the world: Evidence from piaac. *European Economic Review 73*, 103–130.

Hoekstra, M. (2009). The effect of attending the flagship state university on earnings: A discontinuity-based approach. *The review of economics and statistics 91*(4), 717–724.

Instituto Dominicano de Evaluación e Investigación de la Calidad Educativa (IDEICE) (2017). Title of the document. Accessed: 2023-04-28.

Interamerican Development Bank (2023). Latin America and the Caribbean in PISA 2022: How did the Region Perform? Technical report, Interamerican Development Bank.

International, R. (2015). Tout timoun ap li – total (all children reading).

Jackson, C. K. (2018). What do test scores miss? The importance of teacher effects on non–test score outcomes. *Journal of Political Economy 126*(5), 2072–2107.

Jackson, C. K., S. C. Porter, J. Q. Easton, A. Blanchard, and S. Kiguel (2020). School effects on socioemotional development, school-based arrests, and educational attainment. *American Economic Review: Insights 2*(4), 491–508.

Katz, L. F., J. R. Kling, and J. B. Liebman (2001). Moving to opportunity in Boston: Early results of a randomized mobility experiment. *The Quarterly Journal of Economics 116*(2), 607–654.

Kosonen, K. (2005). *Education in local languages: Policy and practice in South-East Asia*. UNESCO Bangkok.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational researcher 49*(4), 241–253.

Kraft, M. A. and G. T. Falken (2020). A blueprint for scaling tutoring across public schools. edworkingpaper no. 20-335. *Annenberg Institute for School Reform at Brown University*.

Kraft, M. A., B. E. Schueler, and G. Falken (2024, August). What impacts should we expect from tutoring at scale? exploring meta-analytic generalizability.

Krueger, A. B. (2003). Economic considerations and class size. *The economic journal 113*(485), F34–F63.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies 76*(3), 1071–1102.

Linden, L., A. Banerjee, and E. Duflo (2003). Computer-assisted learning: Evidence from a randomized experiment. *Poverty Action Lab Paper 5*.

McIntosh, S. and A. Vignoles (2001). Measuring and assessing the impact of basic skills on labour market outcomes. *Oxford Economic Papers 53*(3), 453–481.

Miao, F., W. Holmes, R. Huang, H. Zhang, et al. (2021). *AI and education: A guidance for policymakers*. UNESCO Publishing.

Ministerio de Educación de la República Dominicana (MINERD) (2023, March). Title of the document. Accessed: 2023-04-28.

Mullis, I. V. and M. O. Martin (2017). *TIMSS 2019 Assessment Frameworks.* ERIC.

Muralidharan, K. and A. Singh (2023, 11). Improving schooling productivity through computer-aided instruction:Experimental evidence from Rajasthan*.

Muralidharan, K., A. Singh, and A. J. Ganimian (2019). Disrupting education? experimental evidence on technology-aided instruction in india. *American Economic Review 109*(4), 1426–60.

Nickow, A., P. Oreopoulos, and V. Quan (2020). The impressive effects of tutoring on prek-12 learning: A systematic review and meta-analysis of the experimental evidence.

Nye, B., S. Konstantopoulos, and L. V. Hedges (2004). How large are teacher effects? *Educational evaluation and policy analysis 26*(3), 237–257.

OECD (2023a). *PISA 2022 Results (Volume I).*

OECD (2023b). *PISA 2022 Results (Volume II).*

Piper et al. (2018a). Identifying the essential ingredients to literacy and numeracy improvement: Coaching, teacher professional development, improved books, and teachers' guides. *World Development 106*, 324–336.

Piper et al. (2018b). Scaling up successfully: Lessons from kenya's tusome national literacy program. *Journal of Educational Change 19*(3), 293–321.

Place, K. and P. Gleason (2019). Do charter middle schools improve students' college outcomes? *NCEE Evaluation Brief* (2019-4005).

Reber, S. J., D. Rünger, and M. D. Wong (2023). The effects of charter high schools on academic achievement and college enrollment: Evidence from Los Angeles. *Education Finance and Policy*, 1–19.

Rivkin, S. G., E. A. Hanushek, and J. F. Kain (2005). Teachers, schools, and academic achievement. *Econometrica 73*(2), 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American economic review 94*(2), 247–252.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological monographs: General and applied 80*(1), 1.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, 34–58.

Sass, T. R. (2006, 01). Charter Schools and Student Achievement in Florida. *Education Finance and Policy 1*(1), 91–122.

Sass, T. R., R. W. Zimmer, B. P. Gill, and T. K. Booker (2016). Charter high schools' effects on long-term attainment and earnings. *Journal of Policy Analysis and Management 35*(3), 683–706.

Setren, E. (2021). Targeted vs. general education investments: Evidence from special education and english language learners in boston charter schools. *Journal of Human Resources 56*(4), 1073–1112.

Stern, J. M., M. C. Jukes, J. Cilliers, B. Fleisch, S. Taylor, and N. Mohohlwane (2023). Persistence and emergence of literacy skills: Long-term impacts of an effective early grade reading intervention in south africa. Technical report.

Stiglic, N. and R. M. Viner (2019). Effects of screentime on the health and well-being of children and adolescents: a systematic review of reviews. *BMJ open 9*(1), e023191.

Taylor, E. S. (2015). New technology and teacher productivity. In *CESifo Area Conferences*.

The World Bank (2015, March). Four Things You Need to Know About Education in Haiti.

The World Bank (2021, April). The World Bank in Haiti.

Torres, C. (2022). Classroom management in "no-excuses" charter schools. In *Handbook of Classroom Management*, pp. 152–166. Routledge.

Twenge, J. M. and W. K. Campbell (2018). Associations between screen time and lower psychological well-being among children and adolescents: Evidence from a population-based study. *Preventive medicine reports 12*, 271–283.

UNESCO (2021). Pandemic-related disruptions to schooling and impacts on learning proficiency indicators: A focus on the early grades.

USAID (2019, September). Ghana early grade reading program impact evaluation - 2019 endline report.

Walters, C. R. (2018). The demand for effective charter schools. *Journal of Political Economy 126*(6).

What Works Clearinghouse (2009, August). Success for all. Technical report, U.S. Department of Education, Institute of Education Sciences.

Wong, M. D., K. M. Coller, R. N. Dudovitz, D. P. Kennedy, R. Buddin, M. F. Shapiro, S. H. Kataoka, A. F. Brown, C.-H. Tseng, P. Bergman, et al. (2014). Successful schools and risky behaviors among low-income adolescents. *Pediatrics 134*(2), e389–e396.

World Bank (2017a). *World development report 2018: Learning to realize education's promise*. The World Bank.

World Bank (2017b). *World development report 2018: Learning to realize education's promise*. The World Bank.

World Bank (2020, October). Haiti Human Capital Index 2020.

World Bank (2021). Acting now to protect the human capital of our children: The costs of and response to covid-19 pandemic's impact on the education sector in latin america and the caribbean.

World Bank, UNESCO, UNICEF, USAID, FCDO, and Bill and Melinda Gates Foundation (2022, June 23). The state of global learning poverty: 2022 update.

Young, A. (2019). Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The quarterly journal of economics 134*(2), 557–598.

Zimmerman, S. D. (2014). The returns to college admission for academically marginal students. *Journal of Labor Economics 32*(4), 711–754.

Appendix A: Chapter 1

A.1 Supplemental Tables and Figures

Table A.1: Treatment Effects on Math Overconfidence

	Midline Overconfidence		Endline Overconfidence	
	(1)	(2)	(3)	(4)
CAL Treatment (β_1)	0.018	0.032	-0.002	0.008
	[0.698]	[0.499]	[0.956]	[0.844]
CAL x Tutoring Treatment (β_2)	0.026	0.029	0.042	0.050
	[0.645]	[0.609]	[0.331]	[0.269]
CAL + Tutoring Treatment (β_1 + β_2)	0.044	0.061	0.040	0.058
	[0.351]	[0.203]	[0.289]	[0.132]
Adjusted R^2	-0.00	0.05	-0.00	0.05
Observations	1205	1205	1188	1188
Covariates	No	Yes	No	Yes

Notes: *** p<0.01, ** p<0.05, * p<0.1. Notes: *** p<0.01, ** p<0.05, * p<0.1. The outcome variable takes the value of 1 if the teacher switched classes after program assignment, and 0 otherwise. Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences towards math, self-reported grades for the previous year and parents' employment status. When available, we also control for the baseline level of the outcome variable. The *p*-values shown in brackets represent the probability of obtaining an estimated treatment effect at least as large in absolute value as the one observed in our experiment, under the sharp null hypothesis that the treatment effect is zero for all observations. Randomization inference is based on 2,000 random draws.

Table A.2: Treatment Effects on Student Time Use

	In person classes	Unpaid work (at home)	Unpaid work (out of home)	Paid work (out of school)	Study	Meeting friends	Entertainment	Sports
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
CAL Treatment	-0.060	-0.368	-0.105	-0.132	0.069	-0.072	0.007	0.008
	[0.805]	[0.057]	[0.616]	[0.617]	[0.680]	[0.590]	[0.946]	[0.926]
CAL x Tutoring Treatment	-0.160	0.153	0.018	-0.185	0.071	-0.078	0.041	0.071
	[0.584]	[0.517]	[0.952]	[0.529]	[0.677]	[0.679]	[0.749]	[0.541]
Control Mean	5.99	3.79	2.26	2.74	2.03	2.25	3.33	2.09
Adjusted R^2	0.09	0.08	0.13	0.14	0.02	0.09	0.05	0.22
Observations	1108	1096	1072	1069	1105	1091	1097	1093

Notes: *** p<0.01, ** p<0.05, * p<0.1. Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences towards math, self-reported grades for the previous year and parents' employment status. When available, we also control for the baseline level of the outcome variable. The p-values shown in brackets represent the probability of obtaining an estimated treatment effect at least as large in absolute value as the one observed in our experiment, under the sharp null hypothesis that the treatment effect is zero for all observations. Randomization inference is based on 2,000 random draws.

Table A.3: Treatment Effects on Gender Beliefs

	Midline Gender Beliefs Index		Endline Gender Beliefs Index	
	(1)	(2)	(3)	(4)
CAL Treatment	0.093	0.055	0.037	0.040
	[0.386]	[0.540]	[0.632]	[0.621]
CAL x Tutoring Treatment	-0.107	-0.094	-0.006	-0.027
	[0.415]	[0.344]	[0.950]	[0.796]
Adjusted R^2	0.00	0.13	-0.00	0.02
Observations	1204	1204	1181	1181
Covariates	No	Yes	No	Yes

Notes: *** p<0.01, ** p<0.05, * p<0.1. Notes: *** p<0.01, ** p<0.05, * p<0.1. Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences towards math, self-reported grades for the previous year and parents' employment status. When available, we also control for the baseline level of the outcome variable. The *p*-values shown in brackets represent the probability of obtaining an estimated treatment effect at least as large in absolute value as the one observed in our experiment, under the sharp null hypothesis that the treatment effect is zero for all observations. Randomization inference is based on 2,000 random draws.

Table A.4: Treatment Effects on Midline Math Performance (Lee Bounds)

	Midline Math Test		Endline Math Test	
	Lower Bound	Upper Bound	Lower Bound	Upper Bound
	(1)	(2)	(3)	(4)
CAL Treatment	0.317**	0.086	0.373**	0.233*
	[0.043]	[0.552]	[0.021]	[0.094]
CAL x Tutoring Treatment	-0.175	-0.011	-0.246	-0.241
	[0.395]	[0.961]	[0.223]	[0.168]
Observations	1206	1206	1200	1198

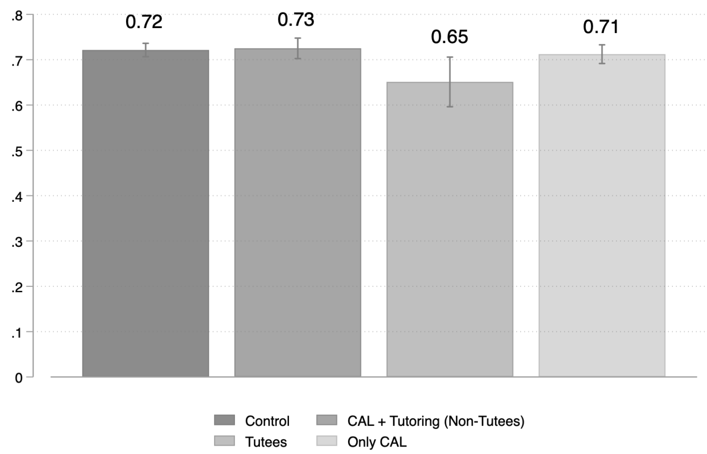
Notes: *** p<0.01, ** p<0.05, * p<0.1. The *p*-values shown in brackets represent the probability of obtaining an estimated treatment effect at least as large in absolute value as the one observed in our experiment, under the sharp null hypothesis that the treatment effect is zero for all observations. Randomization inference is based on 2,000 random draws.

Table A.5: Dose-response of ALEKS Use

	Standardized Math Scores		
	IV estimates	OLS VA (full sample)	OLS VA (treatment group)
	(1)	(2)	(3)
CAL hours	0.0266***	0.0191***	0.0256***
	(0.0056)	(0.0026)	(0.0038)
Baseline math score (SD)	0.1431***	0.1456***	0.1160***
	(0.0098)	(0.0097)	(0.0137)
Observations	1218	1218	617
R-squared	0.3248	0.3296	0.3211
Difference-in-Sargan statistic for exogeneity	0.1304		
Extrapolated estimates of 2.5-hour treatment	0.0665	0.0477	0.0641
Extrapolated estimates 25 hours (1500 minutes)	0.7340	0.5267	0.6767

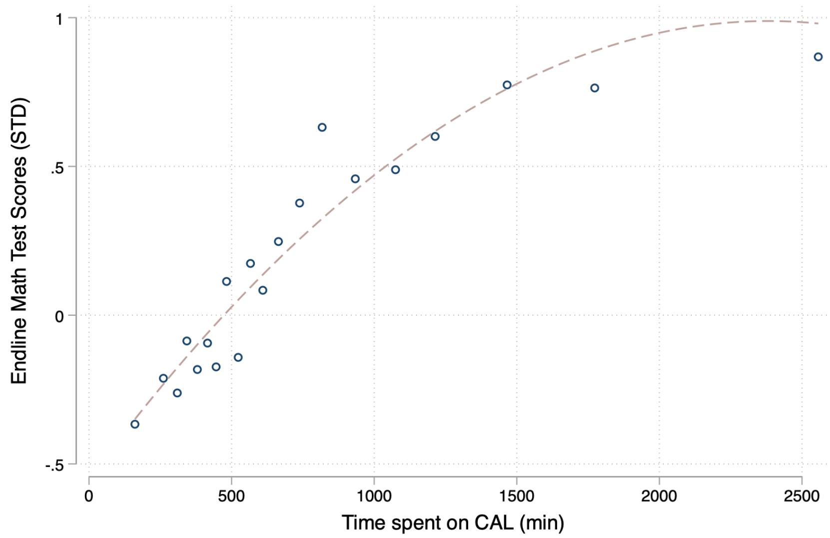
Notes: *** p<0.01, ** p<0.05, * p<0.1. Standard errors in parentheses. Covariates include students' gender, age, baseline math test scores, proxies for confidence and preferences towards math, self-reported grades for the previous year and parents' employment status. Column (1) instruments minutes in ALEKS with the randomized allocation. Column (2) presents OLS value-added models using the full sample. Column (3) presents OLS VA models using only data on the lottery-winners. One hour is equivalent to 60 minutes of platform use.

Figure A.1: Self-Efficacy Index by Performance Groups



Notes: The CAL + Tutoring group excludes tutees.

Figure A.2: Returns to Hours of CAL Use



Notes: This sample is limited to treated students (N = 664).

A.2 Additional Analysis

A.2.1 Power Calculations

To estimate statistical power, we simulate the randomization procedure described in Section 1.3 one thousand times and assume a minimum detectable effect (MDE) of 0.3 standard deviations. Evans and Yuan (2022) find this to be in the 80th percentile of effect sizes of RCTs with learning or access outcomes in low- and middle-income countries. Although this would be above the median of 0.2 for small RCTs, both tutoring and computer-adaptive learning have been identified as two of the most effective interventions to improve learning outcomes. A recent review of tutoring interventions by Nickow et al. (2020) found an overall pooled effect size estimate of 0.37 standard deviations. To the best of our knowledge, there are no meta-studies looking at effects from computer-*adaptive* learning technologies such as the one used in this study.

For this exercise, I use 2019 administrative data on 9th grade students in the same sample of schools selected for this intervention. This data includes students' test scores on a national diagnostic assessment, their sex, age, and a socioeconomic indicator. We also simulate and include a baseline test score variable. These simulations give us a statistical power of approximately 80 percent, which is the standard for randomized controlled trials in the social sciences. We aim to maximize this number by collecting prognostic variables at baseline, including tests scores on a short math assessment and other student characteristics. In Table A.6, we present the results of the simulations using alternative MDEs (0.25, 0.3, 0.35).

MDE	Power
0.25	0.674
0.30	0.799
0.35	0.893

Table A.6: Power Simulations

Appendix B: Chapter 2

B.1 Lottery Details

Table B.1: Massachusetts Charter Schools Eligible for Lottery Study

School (1)	Town (2)	Urban (3)	Grades Served (4)	Lottery Entry Grade(s) (5)	Available Lottery Years (6)
(A) Participating Schools					
Academy of the Pacific Rim Charter School	Boston	Yes	5-12	5,6	2005-2010
Boston Collegiate Charter School	Boston	Yes	5-12	5	2002-2010
Boston Green Academy	Boston	Yes	6-12	9	2011-2014
Boston Preparatory Charter Public School	Boston	Yes	6-12	6	2005-2011, 2014
Cape Cod Lighthouse Charter School	Orleans	No	6-8	6	2007-2010
City on a Hill Charter Public School (Circuit Street)	Boston	Yes	9-12	9	2002, 2004-2014
City on a Hill Charter Public School (Dudley Square)	Boston	Yes	9-12	9	2013-2014
Codman Academy Charter Public School	Boston	Yes	K-12	9	2004, 2008-2014
Edward Brooke Charter School	Boston	Yes	K-12	5	2006-2009
Excel Academy Charter School	Boston	Yes	5-8	5	2008-2010
Four Rivers Charter Public School	Greenfield	No	7-12	7	2003-2012
Francis W Parker Charter Essential School	Devins	No	7-12	7	2006-2011
Global Learning Charter Public School	New Bedford	Yes	5-12	5	2006-2007, 2009
Innovation Academy Charter School	Tyngsboro	No	5-12	5	2007-2010
KIPP Academy Lynn	Lynn	Yes	K-12	5	2005-2009
Marblehead Community Charter Public School	Marblehead	No	4-8	4	2005-2007, 2009
MATCH Charter Public School	Boston	Yes	K-12	6,9	2002-2011
Pioneer Valley Performing Arts Charter Public School	South Hadley	No	7-12	7	2006-2010
Rising Tide Charter Public School	Plymouth	No	5-8	5	2009-2010
Roxbury Preparatory Charter School	Boston	Yes	6-12	6	2002-2011
Salem Academy Charter School	Salem	No	6-12	6	2010-2011
Sturgis Charter Public School	Hyannis	No	9-12	9	2004, 2006, 2008-2011
UP Academy Boston	Boston	Yes	K-8	6	2011
(B) Eligible but Nonparticipating Schools					
Advanced Math and Science Academy Charter School	Marlborough	No	6-12		
Berkshire Arts and Technology Charter Public School	Adams	No	6-12		
Christa McAuliffe Regional Charter Public School	Framingham	Yes	6-8		
Community Charter School of Cambridge	Cambridge	Yes	7-12		
Hampden Charter School of Science	Chicopee	Yes	6-12		
Health Careers Academy Charter School	Boston	Yes	9-12		
New Leadership Charter School	Springfield	Yes	6-12		
Phoenix Charter Academy	Chelsea	Yes	9-12		
Pioneer Charter School of Science	Everett	Yes	7-12		
Sizer School, A North Central Charter Essential School	Fitchburg	Yes	7-12		

Notes: This table includes charter schools that meet the following criteria: 1) admitted students for middle or high-school grades (4-7, 9); 2) has students in the projected high-school classes of 2006-2018; 3) does not serve special populations (such as students a risk of dropping out), and 4) are not closed charter schools. Schools that indicate grade ranges that do not begin at the lottery entry grade expanded grades served after lottery data was collected. Schools in Panel B that are eligible for the lottery study either had fewer applicants than seats available or did not retain usable lottery records. There were 12 closed schools at appropriate grade levels. Urban towns include those which participate in the Massachusetts Urban Superintendents Network: Boston, Brockton, Cambridge, Chelsea, Chicopee, Everett, Fall River, Fitchburg, Framingham, Haverhill, Holyoke, Lawrence, Leominster, Lowell, Lynn, Malden, New Bedford, Pittsfield, Quincy, Revere, Somerville, Springfield, Taunton, and Worcester (Salem joined the network at a later date and thus is categorized as nonurban in this study). MATCH is included in the study as both a middle and a high school.

Table B.2: Covariate Balance

	Urban			Nonurban		
	Fraction of Non-Offered With Outcome (1)	Initial Offer Differential (2)	Waitlist Offer Differential (3)	Fraction of Non-Offered With Outcome (4)	Initial Offer Differential (5)	Waitlist Offer Differential (6)
Female	0.515	-0.003	0.009	0.526	-0.000	0.013
		(0.010)	(0.011)		(0.019)	(0.022)
Asian	0.028	0.004	-0.001	0.025	-0.007	0.015+
		(0.003)	(0.003)		(0.007)	(0.008)
Black	0.519	0.001	-0.000	0.032	-0.002	-0.004
		(0.010)	(0.010)		(0.006)	(0.006)
Latinx	0.280	-0.006	0.009	0.033	0.006	0.000
		(0.009)	(0.009)		(0.007)	(0.008)
Other race	0.042	0.002	-0.008+	0.039	-0.001	-0.014*
		(0.004)	(0.004)		(0.006)	(0.006)
White	0.131	-0.002	-0.000	0.870	0.003	0.002
		(0.006)	(0.006)		(0.012)	(0.013)
Special education	0.196	-0.002	-0.012	0.164	0.004	-0.009
		(0.008)	(0.008)		(0.014)	(0.015)
English learner	0.107	0.000	0.008	0.014	-0.008*	0.004
		(0.006)	(0.007)		(0.004)	(0.005)
Subsidized lunch	0.723	0.008	0.001	0.120	-0.007	0.019
		(0.009)	(0.009)		(0.012)	(0.014)
Baseline MCAS ELA	-0.408	-0.013	0.023	0.436	0.018	-0.049
		(0.021)	(0.022)		(0.032)	(0.037)
Baseline MCAS Math	-0.375	-0.010	0.015	0.322	0.022	-0.015
		(0.020)	(0.021)		(0.036)	(0.041)
p-value		0.951	0.718		0.454	0.364

Notes: This table shows student characteristics and test scores, and differentials between offered and nonoffered charter applicants. The sample is restricted to students enrolled in Massachusetts schools at the time of application in the projected high-school classes of 2006–2018. Columns 1 and 4 show the proportion of non-offered students with a given characteristic. Columns 2, 3, 5, and 6 report coefficients from regressions of the student characteristic on initial and waitlist offer dummies, including controls for risk sets (+ $p<0.10$ * $p<0.05$). N (urban) = 13947, N (nonurban) = 3583.

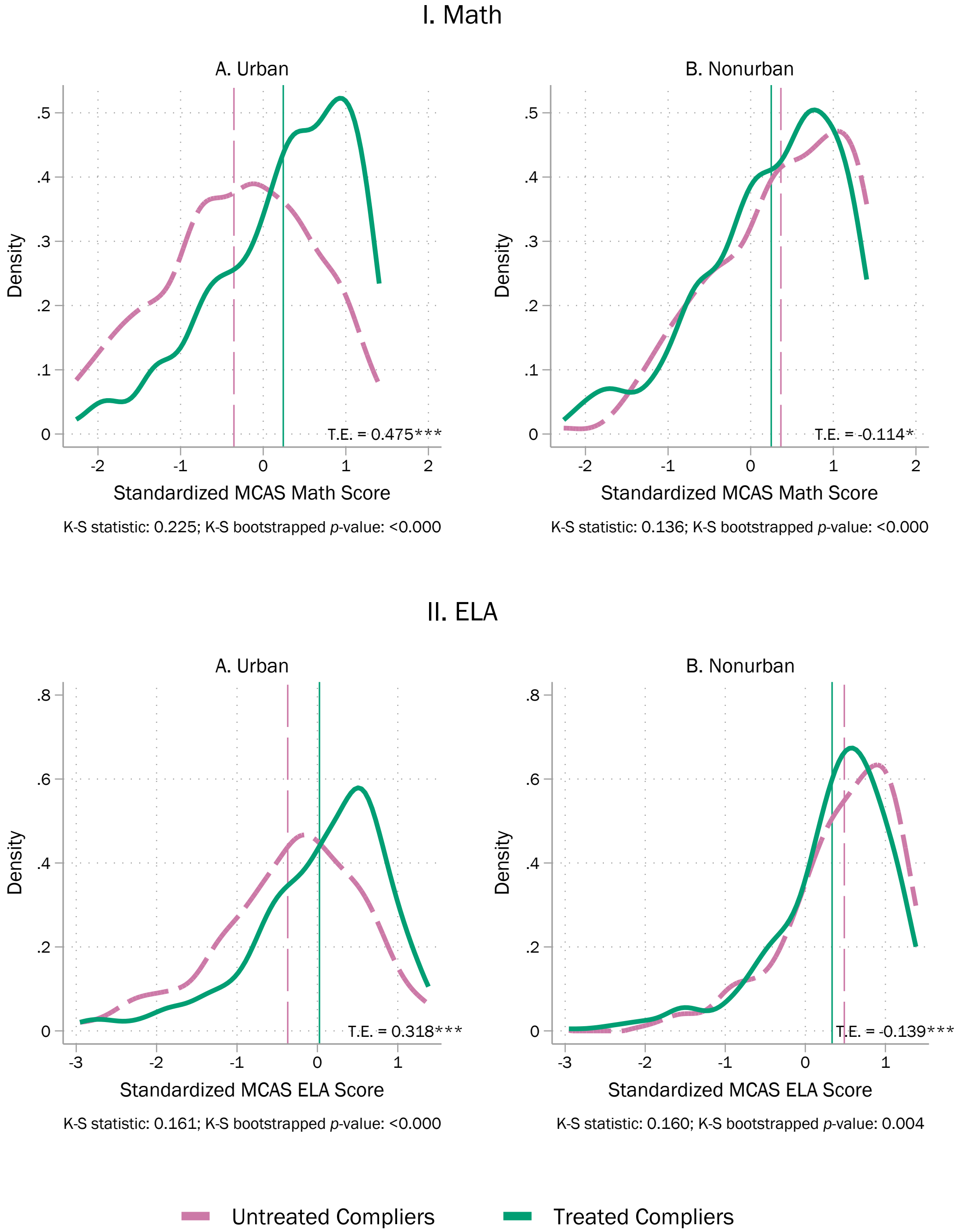
Table B.3: Match Rate to SIMS

Projected HS Class	Non-offered Mean (1)	Initial Offer Differential (2)	Waitlist Offer Differential (3)	Number of Applications (4)
2006	0.986	-0.008	0.008	515
		(0.012)	(0.009)	
2007	0.997	-0.011	-0.033	422
		(0.017)	(0.038)	
2008	0.996	-0.014	0.007	937
		(0.011)	(0.009)	
2009	0.992	0.003	-0.007	1,010
		(0.009)	(0.009)	
2010	0.994	0.000	-0.003	1,332
		(0.009)	(0.010)	
2011	0.996	-0.000	-0.002	1,597
		(0.006)	(0.008)	
2012	0.985	-0.002	0.001	2,142
		(0.006)	(0.005)	
2013	0.991	-0.004	0.001	2,456
		(0.006)	(0.005)	
2014	0.993	0.001	-0.005	2,949
		(0.005)	(0.005)	
2015	0.993	0.000	-0.002	3,746
		(0.005)	(0.004)	
2016	0.992	-0.001	0.001	3,698
		(0.005)	(0.004)	
2017	0.993	-0.000	-0.001	5,226
		(0.004)	(0.003)	
2018	0.995	-0.002	0.001	5,581
		(0.003)	(0.003)	
All cohorts	0.993	-0.002+	-0.000	31,611
		(0.001)	(0.001)	

Notes: This table shows the match between lottery records and the SIMS data by projected high school class. The sample excludes disqualified, late, out-of-area, and sibling applications. Individuals can be in the sample multiple times if they apply to multiple schools. Columns 2 and 3 report coefficients from regressions of the student characteristic on initial and waitlist offer dummies, including controls for risk sets (+ p<0.10 * p<0.05 ** p<0.01 ***p<0.001).

B.2 Additional Results

Figure B.1: Test Score Distributions for Treated and Untreated Compliers



Notes: This figure shows the distribution of test scores for treated and untreated compliers, for MCAS Math and ELA two years after the lottery.

Table B.4: Sample Characteristics and Outcomes

	Urban				Nonurban			
	Noncharter Public Schools (1)	Lottery Applicants (2)	Offered Charter (3)	Not Offered Charter (4)	Noncharter Public Schools (5)	Lottery Applicants (6)	Offered Charter (7)	Not Offered Charter (8)
(A) Baseline characteristics								
Female	0.483	0.518	0.519	0.517	0.491	0.519	0.516	0.526
Asian	0.076	0.030	0.029	0.030	0.039	0.030	0.033	0.025
Black	0.199	0.526	0.538	0.504	0.046	0.026	0.023	0.032
Latinx	0.316	0.284	0.287	0.278	0.062	0.036	0.037	0.033
Other race	0.028	0.041	0.038	0.047	0.017	0.028	0.023	0.039
White	0.380	0.119	0.107	0.142	0.836	0.880	0.884	0.870
Special education	0.187	0.194	0.194	0.195	0.177	0.159	0.156	0.164
English learner	0.179	0.114	0.117	0.109	0.028	0.011	0.010	0.014
Free/reduced price lunch	0.644	0.740	0.752	0.717	0.198	0.128	0.132	0.120
Baseline MCAS ELA	-0.432	-0.418	-0.433	-0.389	0.155	0.417	0.407	0.436
Baseline MCAS Math	-0.426	-0.364	-0.368	-0.356	0.153	0.331	0.336	0.322
(B) Charter school enrollment								
Attend any charter in grades 5-12	0.058	0.427	0.513	0.262	0.052	0.568	0.708	0.284
(C) Academic outcomes								
MCAS Math	-0.359	-0.202	-0.141	-0.311	0.174	0.296	0.297	0.295
MCAS ELA	-0.391	-0.287	-0.248	-0.355	0.189	0.357	0.340	0.394
Took any AP	0.278	0.364	0.378	0.340	0.345	0.313	0.271	0.403
Score 3+ on any AP	0.127	0.102	0.102	0.102	0.244	0.214	0.189	0.266
Took SAT	0.546	0.642	0.638	0.647	0.679	0.771	0.770	0.773
SAT score (1600) (for takers)	94.864	898.188	895.774	902.551	107.148	1125.715	1130.180	1116.332
Graduate high school (4 years)	0.609	0.624	0.608	0.654	0.827	0.821	0.827	0.809
Graduate high school (5 years)	0.656	0.711	0.701	0.731	0.850	0.883	0.880	0.889
Enroll in any college	0.501	0.601	0.604	0.595	0.700	0.799	0.803	0.791
Enroll in 4-year college	0.315	0.446	0.452	0.436	0.563	0.669	0.678	0.651
Enroll in 2-year college	0.217	0.200	0.199	0.201	0.176	0.182	0.176	0.195
Complete any degree	0.257	0.249	0.248	0.251	0.492	0.558	0.566	0.543
Complete BA	0.214	0.214	0.212	0.219	0.449	0.512	0.524	0.490
Complete AA	0.059	0.050	0.052	0.047	0.066	0.070	0.062	0.085
N	275,245	14,191	9,303	4,888	642,470	3,583	2,403	1,180

Notes: This table shows demographic characteristics and outcome means for various samples. Charter attendance includes attendance at non-lottery charters and thus will not match first stage estimates. The sample in Column 1 is restricted to students who attended schools in the state of Massachusetts in 9th grade in the projected high-school classes of 2006–2018. The sample in Column 2 is restricted to charter school applicants enrolled in schools in the state of Massachusetts at the time of application in the projected high-school classes of 2006–2018. The samples in Columns 3 and 4 are further restricted to those offered and not offered a seat at a charter in the lottery, respectively.

Table B.5: The Impact of Charter School Attendance on Advanced Placement

	Urban			Nonurban		
	2SLS (1)	CCM (2)	N (3)	2SLS (4)	CCM (5)	N (6)
(A) Advanced Placement						
Offered AP	0.080***	0.743	13,623	-0.539***	0.786	3,583
	(0.021)			(0.032)		
Number of APs offered	-1.131***	5.727	13,623	-5.659***	8.172	3,583
	(0.269)			(0.411)		
(B) AP by subject						
Offered AP Calculus	0.103***	0.542	13,623	-0.483***	0.702	3,583
	(0.023)			(0.034)		
Offered AP English	-0.136***	0.623	13,623	-0.531***	0.756	3,583
	(0.023)			(0.032)		
Offered AP History	0.021	0.459	13,623	-0.517***	0.730	3,583
	(0.023)			(0.033)		
Offered AP Science	-0.072**	0.511	13,623	-0.490***	0.674	3,583
	(0.023)			(0.032)		
(C) AP scores						
Score 2+ on any AP	0.099***	0.170	13,623	-0.239***	0.411	3,583
	(0.018)			(0.031)		
Score 3+ on any AP	0.045**	0.108	13,623	-0.188***	0.333	3,583
	(0.014)			(0.029)		
Score 4+ on any AP	0.008	0.063	13,623	-0.137***	0.240	3,583
	(0.011)			(0.027)		
Score 5 on any AP	-0.000	0.026	13,623	-0.100***	0.157	3,583
	(0.007)			(0.022)		
(D) Conditional AP scores						
Score 2+ on any AP	0.058+	0.608	3,936	-0.065	0.947	965
	(0.034)			(0.044)		
Score 3+ on any AP	0.008	0.381	3,936	0.014	0.738	965
	(0.034)			(0.064)		
Score 4+ on any AP	-0.038	0.221	3,936	0.002	0.541	965
	(0.028)			(0.077)		
Score 5 on any AP	-0.023	0.089	3,936	-0.071	0.366	965
	(0.018)			(0.074)		

Notes: Each coefficient labeled 2SLS is the instrumental variables estimate of the effect of attending an urban or nonurban charter at any period of time before the outcome listed in the column heading occurred as described in Equation 2.1. Indicator variables for a lottery offer on the day of the lottery (initial offer) and lottery offer from the waitlist (waitlist offer), separately for urban and nonurban charters, are the instruments for charter attendance. The control complier mean is labeled CCM. All regressions control for lottery risk sets and a vector of demographic characteristics including indicators for race, gender, birth year, calendar year, and baseline special education, English learner, and free or reduced price lunch status. The sample is restricted to students enrolled in Massachusetts schools at the time of application in the projected high-school classes of 2006–2018. Robust standard errors in parentheses (+ p<0.10 * p<0.05 ** p<0.01 ***p<0.001). AP outcomes are available for the class of 2007 and later. In the second panel, AP offers are defined based on whether the high school that the student attended offered an AP class. In the third panel, AP scores are conditional on having taken at least one AP.

Table B.6: The Impact of Charter School Attendance on High School Progression

	Urban			Nonurban		
	2SLS (1)	CCM (2)	N (3)	2SLS (4)	CCM (5)	N (6)
(A) On-time grade progression						
10th grade	-0.021 (0.014)	0.914	11,209	-0.002 (0.008)	0.995	3,097
11th grade	-0.045** (0.015)	0.920	10,411	0.006 (0.009)	0.992	3,007
12th grade	-0.036* (0.015)	0.912	10,426	-0.014 (0.011)	0.989	2,966
Repeat 9th or 10th	0.041* (0.019)	0.161	12,001	-0.005 (0.013)	0.027	3,175
(B) High school graduation						
Graduate high school (4 years)	-0.066** (0.023)	0.655	11,983	-0.016 (0.026)	0.815	3,175
Graduate high school (5 years)	-0.018 (0.022)	0.733	11,983	-0.012 (0.024)	0.903	3,175
Graduate high school (6 years)	-0.012 (0.021)	0.787	11,983	-0.017 (0.023)	0.915	3,175
(C) Days attended						
9th grade	0.624 (1.789)	162.716	11,967	1.026 (1.816)	169.753	3,175
10th grade	1.274 (1.597)	161.866	11,199	2.743+ (1.639)	167.366	3,097
11th grade	2.201 (1.769)	156.651	10,401	1.007 (1.709)	167.441	3,007
12th grade	3.722* (1.658)	153.191	10,417	-2.305 (1.618)	160.852	2,966

Notes: Each coefficient labeled 2SLS is the instrumental variables estimate of the effect of attending an urban or nonurban charter at any period of time before the outcome listed in the column heading occurred as described in Equation 2.1. Indicator variables for a lottery offer on the day of the lottery (initial offer) and lottery offer from the waitlist (waitlist offer), separately for urban and nonurban charters, are the instruments for charter attendance. The control complier mean is labeled CCM. All regressions control for lottery risk sets and a vector of demographic characteristics including indicators for race, gender, birth year, calendar year, and baseline special education, English learner, and free or reduced price lunch status. The sample is restricted to students enrolled in Massachusetts schools at the time of application in the projected high-school classes of 2006–2018. Robust standard errors in parentheses (+ p<0.10 * p<0.05 ** p<0.01 ***p<0.001).

Table B.7: The Impact of Charter School Attendance on College Enrollment

Year after Projected High School Graduation	Any College 2SLS (1)	Any College CCM (2)	4 Year College 2SLS (3)	4 Year College CCM (4)	2 Year College 2SLS (5)	2 Year College CCM (6)	N (7)
(A) 1st year							
Urban	0.024	0.518	0.063**	0.390	-0.039**	0.127	13,281
	(0.023)		(0.023)		(0.015)		
Nonurban	0.022	0.665	0.094**	0.529	-0.072**	0.135	3,414
	(0.031)		(0.033)		(0.022)		
(B) 2nd year							
Urban	0.077***	0.495	0.079***	0.367	-0.002	0.127	13,281
	(0.023)		(0.022)		(0.016)		
Nonurban	0.063+	0.684	0.098**	0.557	-0.035	0.127	3,414
	(0.032)		(0.034)		(0.023)		
(C) 3rd year							
Urban	0.073**	0.439	0.064**	0.335	0.010	0.103	13,281
	(0.023)		(0.022)		(0.015)		
Nonurban	0.095**	0.631	0.099**	0.545	-0.004	0.086	3,414
	(0.034)		(0.035)		(0.020)		
(D) 4th year							
Urban	0.068**	0.396	0.053*	0.310	0.015	0.085	13,281
	(0.023)		(0.021)		(0.014)		
Nonurban	0.121***	0.572	0.125***	0.510	-0.004	0.062	3,414
	(0.035)		(0.035)		(0.017)		
(E) 5th year							
Urban	0.015	0.245	0.023	0.179	-0.010	0.066	13,281
	(0.020)		(0.018)		(0.012)		
Nonurban	0.029	0.299	0.041	0.257	-0.013	0.043	3,414
	(0.034)		(0.033)		(0.013)		
(F) 6th year							
Urban	0.049**	0.132	0.043**	0.087	0.006	0.044	11,608
	(0.019)		(0.016)		(0.011)		
Nonurban	0.040	0.151	0.048+	0.130	-0.009	0.023	3,158
	(0.028)		(0.027)		(0.011)		

Notes: Each coefficient labeled 2SLS is the instrumental variables estimate of the effect of attending an urban or nonurban charter at any period of time before the outcome listed in the column heading occurred as described in Equation 2.1. Indicator variables for a lottery offer on the day of the lottery (initial offer) and lottery offer from the waitlist (waitlist offer), separately for urban and nonurban charters, are the instruments for charter attendance. The control complier mean is labeled CCM. All regressions control for lottery risk sets and a vector of demographic characteristics including indicators for race, gender, birth year, calendar year, and baseline special education, English learner, and free or reduced price lunch status. The sample is restricted to students enrolled in Massachusetts schools at the time of application in the projected high-school classes of 2006–2018. Robust standard errors in parentheses (+ p<0.10 * p<0.05 ** p<0.01 ***p<0.001).

Table B.8: The Impact of Charter School Attendance on College Degrees

	Any Degree		B.A.		A.A.		
Year after Projected High School Graduation	2SLS (1)	CCM (2)	2SLS (3)	CCM (4)	2SLS (5)	CCM (6)	N (7)
(A) 4th year							
Urban	0.037*	0.133	0.031*	0.116	0.008	0.022	13,281
	(0.016)		(0.015)		(0.008)		
Nonurban	0.061+	0.364	0.074*	0.308	-0.027	0.074	3,414
	(0.035)		(0.034)		(0.017)		
(B) 5th year							
Urban	0.038*	0.211	0.033+	0.188	0.003	0.037	13,281
	(0.019)		(0.018)		(0.009)		
Nonurban	0.105**	0.483	0.115**	0.432	-0.031+	0.085	3,414
	(0.036)		(0.036)		(0.018)		
(C) 6th year							
Urban	0.046*	0.240	0.042*	0.216	0.009	0.041	11,608
	(0.022)		(0.021)		(0.011)		
Nonurban	0.112**	0.519	0.119**	0.472	-0.030	0.093	3,158
	(0.037)		(0.037)		(0.020)		

Notes: Each coefficient labeled 2SLS is the instrumental variables estimate of the effect of attending an urban or nonurban charter at any period of time before the outcome listed in the column heading occurred as described in Equation 2.1. Indicator variables for a lottery offer on the day of the lottery (initial offer) and lottery offer from the waitlist (waitlist offer), separately for urban and nonurban charters, are the instruments for charter attendance. The control complier mean is labeled CCM. All regressions control for lottery risk sets and a vector of demographic characteristics including indicators for race, gender, birth year, calendar year, and baseline special education, English learner, and free or reduced price lunch status. The sample is restricted to students enrolled in Massachusetts schools at the time of application in the projected high-school classes of 2006–2018. Robust standard errors in parentheses (+ p<0.10 * p<0.05 ** p<0.01 ***p<0.001). Students can obtain both a BA and an AA, so the coefficient for any degree will not be the sum of the BA and AA coefficients.

Table B.9: The Impact of Charter School Attendance on Key Variables, Alternative Specifications

	Urban			Nonurban		
	Math MCAS (1)	4-Year College Enrollment (2)	4-Year College Graduation (3)	Math MCAS (4)	4-year College Enrollment (5)	4-year College Graduation (6)
Main Specification	0.475***	0.079***	0.042*	-0.114+	0.098**	0.119**
	(0.040)	(0.022)	(0.021)	(0.059)	(0.034)	(0.037)
N	10,706	13,281	11,608	3,198	3,414	3,158
Initial offer only	0.424***	0.062*	0.036	-0.086	0.103*	0.111*
	(0.052)	(0.031)	(0.029)	(0.075)	(0.044)	(0.046)
N	10,706	13,281	11,608	3,198	3,414	3,158
Baseline test scores	0.483***	0.076**	0.039+	-0.091	0.104**	0.127***
	(0.041)	(0.023)	(0.022)	(0.062)	(0.036)	(0.039)
N	9,923	11,996	10,456	2,714	2,869	2,665
No covariates	0.500***	0.085***	0.044*	-0.100	0.097**	0.121**
	(0.044)	(0.023)	(0.021)	(0.064)	(0.035)	(0.038)
N	10,706	13,281	11,608	3,198	3,414	3,158

Notes: The first row of the table repeats the main specification reported in the other tables, see Tables 2.2 and 2.3 for details. Four-year college enrollment is enrollment within 2 years of projected high school graduation. Each subsequent row shows an alternative specification. The row labeled Initial offer only uses only admissions offers on the day of the charter school lottery as an instrument for charter attendance (excluding the waitlist offer). The row labeled Baseline test scores uses only students for which baseline math MCAS scores are available. The row labeled No covariates excludes all demographic variables from the regression.Robust standard errors in parentheses (+ $p<0.10$ * $p<0.05$ ** $p<0.01$ ***$p<0.001$).

Appendix C: Chapter 3

C.1 Supplemental Tables and Figures

Table C.1: EGRA Components

Component	Early reading skill	Skill demonstrated by students' ability to:
1. Letter name knowledge	Letter recognition	- Provide the name of upper- and lowercase letters in random order
2. Phonemic awareness	Phonemic awareness	- Segment words into phonemes - Identify the initial sounds in different words
3. Letter sound knowledge	Phonics	- Provide the sound of upper- and lowercase letters distributed in random order
4. Familiar word reading	Word reading	- Read simple and common one- and two-syllable words
5. Unfamiliar nonword reading	Alphabetic principle	- Make grapheme-phoneme correspondences (GPCs) through the reading of simple nonsense words
6. Oral reading fluency with comprehension	Oral reading fluency, reading comprehension	- Read a text with accuracy, with little effort, and at a sufficient rate - Respond correctly to different types of questions, including literal and inferential questions about the text they have read
7. Listening comprehension	Listening comprehension	- Respond correctly to different types of questions including literal and inferential questions about the text the enumerator reads to them
8. Dictation	Alphabetic principle	- Write, spell, and use grammar properly through a dictation exercise

Source: RTI International (2009). Early Grade Reading Assessment Toolkit.

Table C.2: Variation in the Number of Visits After Initial Diagnosis

	Before		After	
	Supervisor visits	Coach visits	Supervisor visits	Coach visits
Full School	1 per month	1 at least every fifteen days	1 at least per month	2 at least every fifteen days
Light Schools	1 per month	0	1 per month	1 per month (in schools with more difficulties)*

Source: Programme M'ap Li Net Ale Rapport Finale (CEEC). *For coaching visits at Light schools, schools whose principals had not attended initial training sessions were prioritized.

Table C.3: Summary of Mobilization Activities by Geographical Department

Department	Meetings	Activities	Schools Involved	Students
Grand-Anse	65	13	45	1295
Nippes	93	36	128	2229
Northwest	56	18	12	1261
West	82	29	47	985

Source: Programme M'ap Li Net Ale Rapport Finale (CEEC).

Table C.4: Socioeconomic Characteristics for Students at Baseline

Characteristic	Share
Age	7.25
Male	0.54
The pupil lives with his parents	
Dad	0.74
Mom	0.84
Both	0.68
Student attended preschool	0.76
The student ate breakfast before going to school	0.71
Student's school equipment	
Pencil	0.74
Pen	0.05
Chalk	0.05
Slate	0.06
Notebook	0.76
Family property and utilities	
Electricity	0.33
Radio	0.49
Television	0.20
Fridge	0.11
Computer	0.05
Bicycle	0.11
Motorbike	0.12
Car	0.06

Source: Baseline implementation report.

Table C.5: Schools and Students by Treatment Group at Midline and Endline

Department	Schools				Students			
	Control (1)	Full (2)	Light (3)	Total (4)	Control (5)	Full (6)	Light (7)	Total (8)
Panel A. Midline Sample								
Grande-Anse	34	30	32	96	1142	1037	1135	3314
Nippes	62	66	63	191	1527	1608	1667	4802
Ouest	24	28	25	77	648	743	711	2102
Nord-Ouest	24	26	26	76	811	853	744	2408
Total	144	150	146	440	4128	4241	4257	12626
Panel B. Endline Sample								
Grande-Anse	34	30	32	96	1119	992	1066	3177
Nippes	60	66	63	189	1539	1704	1766	5009
Ouest	24	27	23	74	651	710	656	2017
Nord-Ouest	24	25	76	27	754	812	751	2317
Total	142	148	145	435	4063	4218	4239	12520

Source: Author's own estimates. Data covers two cohorts, students in midline and endline samples.

Table C.6: Additional EGRA Results at Baseline: Fluency Scores

Task	Control (1)	Light (2)	Full (3)	Difference L-C (4)	Difference F-C (5)	Difference F-L (6)
Letter recognition	10.48	11.86	10.07	1.38	-0.41	-1.79
Sound recognition	7.68	8.85	7.03	1.17	-0.65	-1.82
Familiar word reading	4.56	5.29	4.01	0.73	-0.55	-1.28
Reading invented words	3.40	4.51	3.10	1.10	-0.30	-1.41
Oral reading fluency	6.45	6.14	5.03	-0.31	-1.42	-1.10

Notes: *p<0.10, **p<0.05, ***p<0.01. Standard errors are clustered at the school level. Fluency scores are respectively computed as: i) total number of letters out of 100 for which the student correctly identifies the name in one minute; ii) number of letters sounds correctly identified in one minute; iii) number of familiar words correctly read in one minute; iv) number of coined words correctly read in one minute; and v) number of words correctly read in one minute from a short text. Because response time was not measured for these outcomes, we do not estimate fluency scores for the following tasks: phonemic awareness, reading comprehension, listening comprehension and word dictation.

Table C.7: Additional EGRA Results at Baseline: Percent with Zero Scores

Task	Control (1)	Light (2)	Full (3)	Difference L-C (4)	Difference F-C (5)	Difference F-L (6)
Letter recognition	21.02	24.07	24.95	3.05	3.93	0.88
Phonemic awareness	78.44	73.56	70.21	-4.87*	-8.23***	-3.36
Sound recognition	30.89	35.28	33.22	4.39	2.33	-2.06
Familiar word reading	58.65	60.37	62.59	1.72	3.94	2.23
Reading invented words	77.36	76.66	78.48	-0.70	1.12	1.82
Oral reading fluency	60.70	61.77	62.78	1.07	2.08	1.01
Reading comprehension	82.86	83.57	81.20	0.71	-1.65	-2.36
Listening comprehension	24.85	28.71	23.87	3.86*	-0.98	-4.84**
Word dictation	37.47	39.73	42.01	2.26	4.54	2.28
Overall composite score	5.98	8.46	6.20	2.47**	0.22	-2.26*

Notes: *p<0.10, **p<0.05, ***p<0.01. Standard errors are clustered at the school level. Estimates for each task represent the share of students with a zero score.

Table C.8: School Characteristics at Baseline

Characteristic	Control (1)	Light (2)	Full (3)	Diff. L-C (4)	Diff. F-C (5)	Diff. F-L (6)
Panel A. School Characteristics						
School has preschool	0.64	0.67	0.71	0.03	0.07	0.04
School has 3rd cycle of elementary	0.30	0.29	0.27	-0.01	-0.03	-0.02
Number of first grade classrooms	1.09	1.06	1.23	-0.03	0.14**	0.17
Number of second year rooms	1.06	1.11	1.08	0.05	0.02	-0.03
Vacation type						
Simple vacation	0.87	0.91	0.86	0.04	-0.01	-0.05
Double shift	0.12	0.08	0.14	-0.04	0.02	0.06
Number of hours of single shift	5.05	5.04	5.00	-0.01	-0.05	-0.04
School benefits from canteen program	0.24	0.20	0.23	-0.04	-0.01	0.03
School benefits from deworming program	0.52	0.41	0.41	-0.11*	-0.11	0.00
Panel B. School Infrastructure						
School has a school latrine	0.92	0.93	0.92	0.01	0.00	-0.01
School has drinking water	0.59	0.55	0.55	-0.04	-0.04	0.00
School has a library	0.23	0.24	0.22	0.01	-0.01	-0.02
School building type						
A church	0.05	0.08	0.13	0.03**	0.08**	0.05
A building built for a school	0.77	0.78	0.75	0.01	-0.02	-0.03
A temporary building or hangar	0.08	0.08	0.05	0.00	-0.03	-0.03
Roof of the building						
Concrete	0.20	0.22	0.16	0.02	-0.04	-0.06
Wood	0.01	0.03	0.01	0.02	0.00	-0.02
Straw	0.01	0.00	0.01	-0.01	0.00	0.01
Sheet metal	0.75	0.74	0.79	-0.01	0.04	0.05
Building wall						
Concrete	0.91	0.85	0.88	-0.06	-0.03	0.03
Wood	0.05	0.08	0.04	0.03	-0.01	-0.04
Straw	0.01	0.01	0.01	0.00	0.00	0.00
Sheet metal	0.02	0.07	0.04	0.05**	0.02	-0.03**

Notes: *p<0.10, **p<0.05, ***p<0.01. This table is copied from a baseline implementation report.

Table C.9: School Management Characteristics at Baseline

Characteristic	Control (1)	Light (2)	Full (3)	Diff. L-C (4)	Diff. F-C (5)	Diff. F-L (6)
School has a parents' committee	0.91	0.85	0.89	-0.06	-0.02	0.04
Number of PC meetings since start of school year	1.97	2.29	2.13	0.32	0.16	-0.16*
School has a school management council	0.90	0.91	0.85	0.01	-0.05	-0.06
Teaching staff						
Number of women in first grade	0.91	0.87	0.90	-0.04	-0.01	0.03
Number of women in second year	0.59	0.63	0.44	0.04*	-0.15***	-0.19
Availability of teachers' attendance register						
Yes, seen an up-to-date register	0.74	0.74	0.72	0.00	-0.02	-0.02
Yes, seen an outdated register	0.12	0.13	0.12	0.01	0.00	-0.01
Yes, not seen	0.10	0.08	0.08	-0.02	-0.02	0.00
No	0.04	0.05	0.08	0.01	0.04	0.03
Availability of student attendance register						
Yes, seen an up-to-date register	0.84	0.84	0.81	0.00	-0.03	-0.03
Yes, seen an outdated register	0.09	0.08	0.09	-0.01	0.00	0.01
Yes, not seen	0.06	0.08	0.09	0.02	0.03	0.01
No	0.00	0.01	0.01	0.01	0.01	0.00
School allows to bring home library books	0.12	0.15	0.13	0.03	0.01	-0.02

Source: Baseline implementation report.

Table C.10: EGRA Standardized Results at Baseline

Task	Control (1)	Light (2)	Full (3)	Difference L-C (4)	Difference F-C (5)	Difference F-L (6)
Letter recognition	0.02	0.01	-0.02	-0.01	-0.04	-0.03
Phonemic awareness	-0.07	0.01	0.06	0.08	0.13*	0.05
Sound recognition	-0.01	0.04	-0.03	0.05	-0.02	-0.06
Familiar word reading	-0.01	0.04	-0.03	0.05	-0.03	-0.07
Reading invented words	-0.02	0.05	-0.03	0.07	-0.02	-0.08
Oral reading fluency	0.00	0.04	-0.04	0.03	-0.04	-0.07
Reading comprehension	-0.01	-0.03	0.03	-0.02	0.04	0.06
Listening comprehension	0.07	-0.10	0.04	-0.17***	-0.04	0.13***
Dictated	0.05	0.00	-0.04	-0.05	-0.09	-0.04
Overall composite score	0.00	0.00	0.00	0.00	-0.01	0.00

Notes: *p<0.10, **p<0.05, ***p<0.01. Standard errors are clustered at the school level.

Table C.11: EGRA Standardized Results at Midline

Task	Control (1)	Light (2)	Full (3)	Difference L-C (4)	Difference F-C (5)	Difference F-L (6)
Letter recognition	0.40	0.35	0.37	-0.05	-0.03	0.02
Phonemic awareness	-0.03	0.25	0.32	0.28***	0.35***	0.06
Sound recognition	0.16	0.33	0.32	0.16**	0.15**	-0.01
Familiar word reading	0.26	0.31	0.30	0.05	0.05	-0.01
Reading invented words	0.22	0.27	0.26	0.05	0.05	0.00
Oral reading fluency	0.29	0.33	0.30	0.03	0.01	-0.03
Reading comprehension	0.21	0.26	0.23	0.06	0.02	-0.04
Listening comprehension	0.27	0.24	0.34	-0.03	0.07*	0.10**
Dictation	0.44	0.41	0.40	-0.03	-0.05	-0.01
Overall composite score	0.31	0.40	0.42	0.08	0.11	0.02

Notes: *p<0.10, **p<0.05, ***p<0.01. Standard errors are clustered at the school level.

Table C.12: EGRA Standardized Results at Endline

Task	Control (1)	Light (2)	Full (3)	Difference L-C (4)	Difference F-C (5)	Difference F-L (6)
Letter recognition	0.35	0.36	0.31	0.01	-0.04	-0.05
Phonemic awareness	-0.32	0.44	0.49	0.76***	0.80***	0.05
Sound recognition	-0.10	0.44	0.43	0.54***	0.53***	-0.01
Familiar word reading	0.17	0.35	0.32	0.18**	0.15*	-0.03
Reading invented words	0.17	0.31	0.27	0.14	0.10	-0.04
Oral reading fluency	0.17	0.35	0.31	0.18**	0.14	-0.04
Reading comprehension	0.15	0.33	0.29	0.18**	0.14*	-0.04
Listening comprehension	0.17	0.23	0.28	0.06	0.11**	0.05
Dictated	0.31	0.45	0.39	0.14**	0.08	-0.06
Overall composite score	0.12	0.47	0.45	0.34***	0.33***	-0.02

Notes: *p<0.10, **p<0.05, ***p<0.01. Standard errors are clustered at the school level.

Table C.13: Implementation Status at Midline

Variable	Light (1)	Full (2)	Difference L-F (3)
Share of schools participating in the program	95.35	94.30	1.05
Days of training	4.77	5.71	-0.94
Number of staff trained	2.07	1.99	0.08
Share of schools where all 1st-year teachers used the methodology	90.60	98.00	-7.40
Reasons for not using the methodology			
Salary issues related to delays	100.00	0.00	100.00
Using other methods	0.00	30.18	-30.18
Other	0.00	69.82	-69.82
School received teaching material	96.35	82.71	13.64
School staff trained in the methodology			
Pedagogical principal was trained	85.87	65.44	20.43
Administrative principal was trained	12.81	26.40	-13.59
First-year teachers were trained	98.14	91.53	6.61
Received school supplies			
Students received books	0.67	0.64	0.03
Students received notebooks	0.67	0.57	0.10
Students benefited from the methodology	0.96	0.93	0.03
Visits from external supervisors per month	1.59	3.20	-1.61
Share of schools that participated in extracurricular activities	52.34	62.08	-9.74
Share of schools where teachers reported continuation	97.36	66.76	30.60
Observations	3,758	3,215	6,973
Number of schools	146	150	296

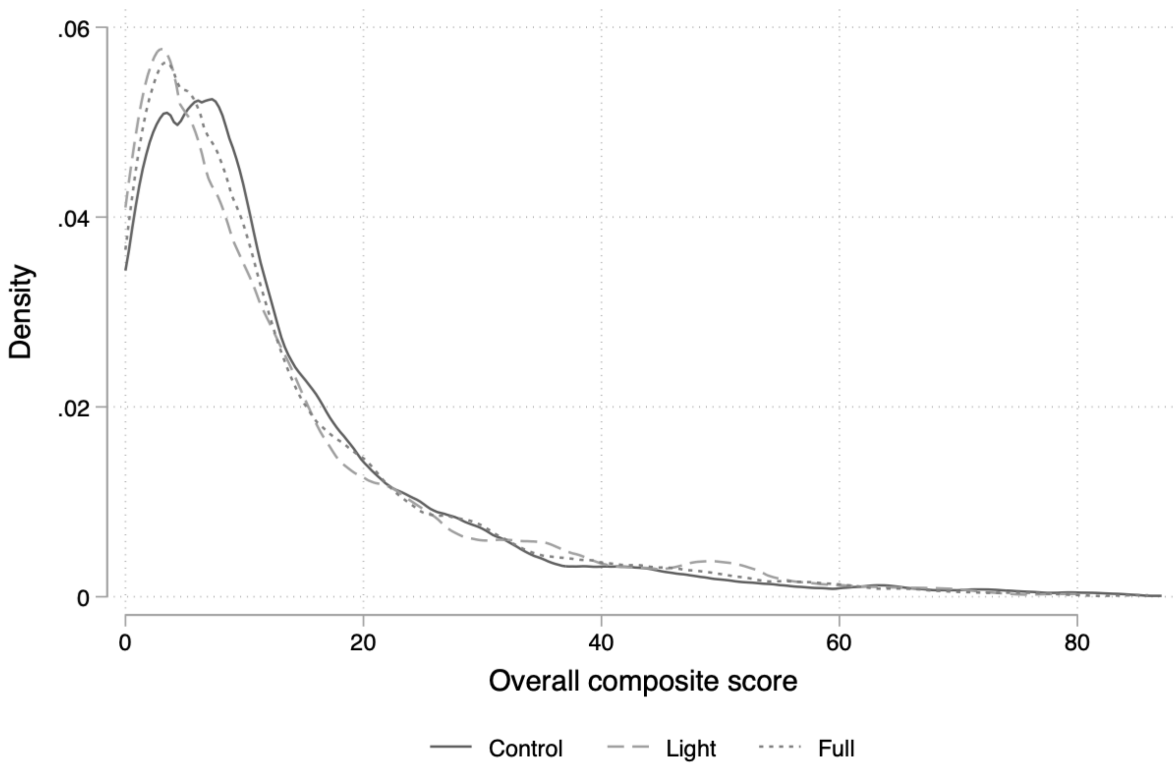
Notes: *** p<0.01, ** p<0.05, * p<0.1. Standard errors are clustered at the school level.

Table C.14: Implementation Status at Endline

Variable	Light (1)	Full (2)	Difference L-F (3)
Share of schools participating in the program	95.20	95.23	-0.03
Days of training	4.87	5.59	-0.72
Number of staff trained	2.88	2.81	0.07
Share of schools where all 1st-year teachers used the methodology	84.85	83.93	0.92
Reasons for not using the methodology			
Salary issues related to delays	11.23	11.79	-0.56
Using other methods	5.15	2.24	2.91
Other	83.62	85.97	-2.35
School received teaching material	93.05	91.67	1.38
School staff trained in the methodology			
Pedagogical principal was trained	69.02	75.28	-6.26
Administrative principal was trained	27.62	20.66	6.96
First-year teachers were trained	98.86	98.19	0.67
Second-year teachers were trained	96.77	94.53	2.24
Received school supplies			
Students received books	0.81	0.76	0.05
Students received notebooks	0.81	0.81	0.00
Visits from external supervisors per month			
1st Year	1.59	1.18	0.41
2nd Year	1.54	1.18	0.36
Share of schools that participated in extracurricular activities	19.90	25.34	-5.44
Share of schools where teachers reported continuation	96.90	94.27	2.63
Observations	4,239	4,218	8,457
Number of schools	145	148	293

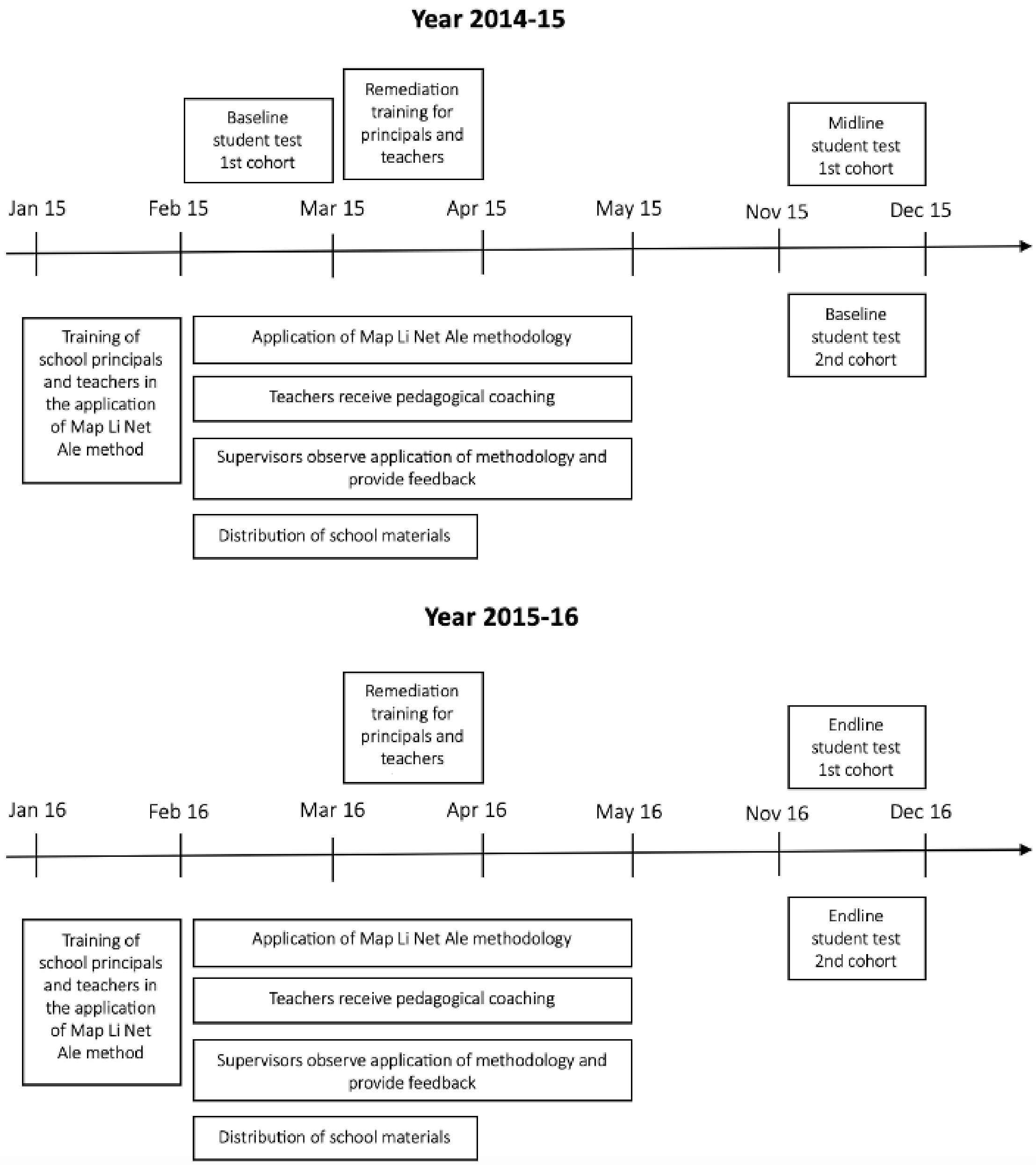
Notes: *** p<0.01, ** p<0.05, * p< 0.1. Standard errors are clustered at the school level.

Figure C.1: Kernel Distribution of Composite Score at Baseline by Treatment Group



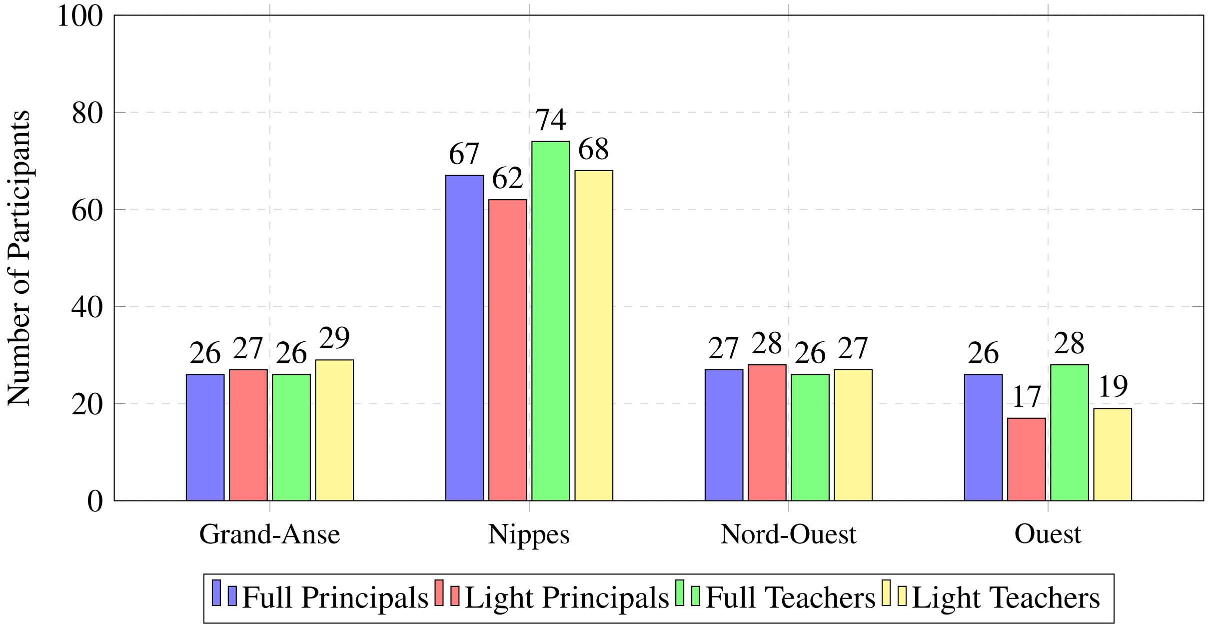
Notes: We conducted a two-sample Kolmogorov-Smirnov test for equality of distribution functions to check for differences across treatment groups. No statistically significant differences were found between the Light and Full groups (combined p-value 0.173), and Control and Full groups (combined p-value 0.058) at the 5 percent significance level. We found statistically significant differences between the Control and Light groups (combined p-value 0.000). This test omits the nested structure of the data.

Figure C.2: Full Timeline of Program Implementation and Data Collection



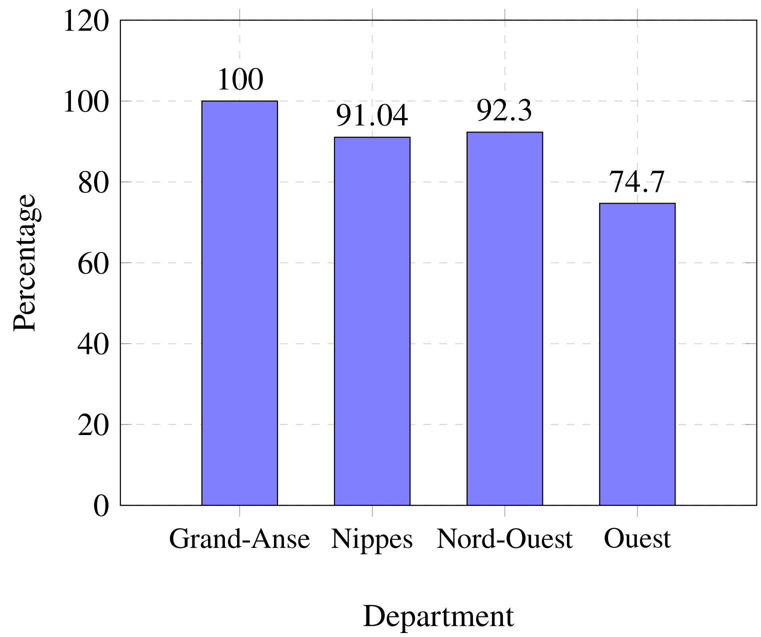
Notes: The first cohort is composed of students enrolled in the 1st grade in the 2014-15 school year. The second cohort includes students enrolled in the 1st grade in the 2015-16 school year.

Figure C.3: Participation in Training of Principals and Teachers by Type of School



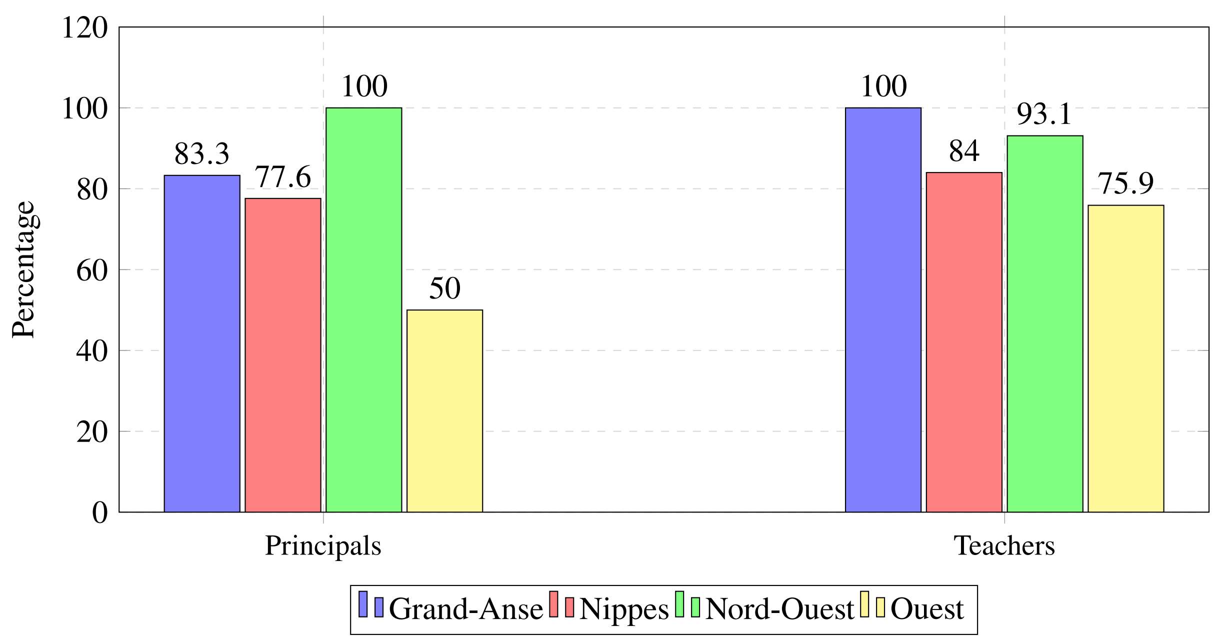
Notes: This graph shows the number of principals and teachers participating in training sessions, categorized by department and type of school (Full or Light).

Figure C.4: Training Participation Rates by Geographical Department



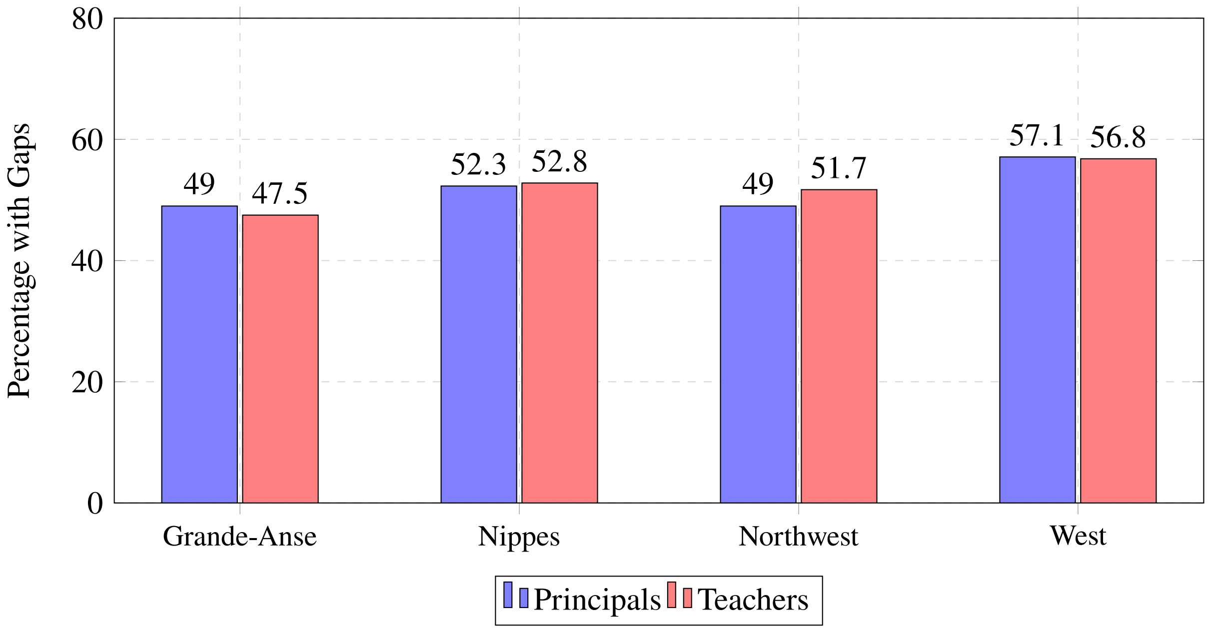
Notes: This graph shows school participation rates in training sessions by geographical department.

Figure C.5: Training Participation Rates by Participant Category and Geographical Department



Notes: This graph shows attendance rates by participant category (principals and teachers) across different geographical departments. Source: Programme M'ap Li Net Ale Rapport Finale (CEEC).

Figure C.6: Percentage of Principals and Teachers with Identified Gaps by Department



Notes: This graph displays the percentage of principals and teachers with identified gaps across geographical departments. Source: Programme M'ap Li Net Ale Rapport Finale (CEEC).