Nurturing Childhood Curiosity to Enhance Learning:

Evidence from a Randomized Pedagogical Intervention

By Sule Alan and Ipek Mumcu*

We evaluate a pedagogical intervention aiming at improving learning in elementary school children by fostering their curiosity. We test the effectiveness of the pedagogy using achievement scores and a novel measure of curiosity. The latter involves creating a sense of information deprivation and quantifying the urge to acquire information and retention ability. The intervention increases curiosity, knowledge retention, and science test scores, with the effects persisting into middle school years. It also leads to more information sharing and peer learning in the classroom. The evidence can help design better pedagogical tools to increase pupil engagement and the quality of learning.

JEL: I24, I26

Keywords: Learning crisis, Curiosity, Deep Learning, Pedagogy, Achievement

Today, more children than ever enroll in primary and post-primary education in the developing world. Despite this progress, the quality of education remains low. Millions of children in developing countries leave school without the necessary foundational skills to help them achieve their potential and lead productive lives.[1] Low teacher quality, overcrowded classrooms, and inadequate levels of school inputs such as poorly designed curricula and insufficient teaching materials are among the many factors contributing to low learning outcomes (Glewwe and Muralidharan, 2016; Glewwe, Lambert and Chen, 2020). Recent research highlights the role of pedagogy as a potentially effective policy tool to combat poor education quality (World Bank, FCDO and BE2, 2020; Brown et al., 2022). While there is no consensus on what constitutes good pedagogy, teaching practices that respond to the needs of students at all levels, build on their individual

* Alan: European University Institute, Italy, and Bilkent University, Turkey, salancrossley@gmail.com. Mumcu: University of Exeter, United Kingdom, I.Mumcu@exeter.ac.uk. We are grateful to J-PAL Post-Primary Education Initiative and ING Bank Turkey for funding this study. We thank seminar participants at EUI, LSE, Bologna ESA 2022, the University of Michigan, the University of Toronto and Cornell University for their valuable comments. We thank Enes Duysak, Mert Gumren, Canan Guner, Elif Kubilay, Ozge Seyrek, Fatima Silpagar and Melek Celik for wonderful field assistance and Selin Iplikci for excellent research assistance. The study has ethics approval from the University of Essex Ethics Board (IRB#: IRB00000522). Study 1 AEA Registry: AEARCTR-0003957 (Alan and Mumcu, 2023). Study 2 AEA Registry: AEARCTR-0008629 (Alan, 2021).

[1] According to 2017 Annual Status of Education Report for India, about 25% aged 14-18 fail to read basic text fluently in their language, 57% struggle with division (three digits by one digit) (ASER, 2018). Results from similar tests in Pakistan and East Africa paint a similar picture. PISA and TIMMS results highlight large learning gaps between the developing and the developed world (Gust, Hanushek and Woessmann, 2022). The recent study by Singh, Romero and Muralidharan (2022) documents the further damage done to learning by the Covid-19 pandemic.

strengths, and encourage them to learn through experimentation are likely to be effective.[2] Unfortunately, most traditional instruction techniques lack these features. They ignore heterogeneous learning paths, compel students to be passive listeners, and prevent the development of an active and inquisitive mind (Blanchard, Southerland and Granger, 2009; Granger et al., 2012; Terrenghi et al., 2019; Ashraf, Banerjee and Nourani, 2021).

In this paper, we evaluate the effectiveness of a pedagogical program that aims to nurture children's curiosity and improve learning outcomes. The program targets teachers' everyday teaching practices, encouraging them to be more creative in preparing children for deep learning experiences before introducing new subjects. Grounded in recent insights on the neural mechanisms of human curiosity and its connection to deep learning, the program provides teachers with knowledge on how learning is enhanced in the brain when the urge to acquire knowledge is stimulated. This knowledge equips teachers with strategies to improve the quality of learning in their classrooms. For this, treated teachers first underwent seminar sessions that explained the mechanisms of learning and the formation of long-term memories. Then, they studied interactively a pedagogical toolkit containing various visual and reading materials prepared for them. The toolkit contains innovative ideas for creating teachable moments and holding students' undivided attention before introducing new and complex curricular topics. While it may be utilized in teaching any curricular topic, in its current form, the pedagogy is more relevant for science teaching as the toolkit primarily targets scientific curiosity. Teachers were asked to practice the prescribed pedagogy throughout an academic year.

Curiosity, a fundamental component of human cognition, is considered a critical driver of success in most aspects of life. Berlyne (1954) and Loewenstein (1994) provide a theoretical framework for epistemic curiosity, described as "desire for knowledge".[3] Cognitive psychology associates curiosity with achievement in many domains ranging from education to health and overall life satisfaction (Chamorro-Premuzic and Furnham, 2006; Kashdan and Silvia, 2009; von Stumm, Hell and Chamorro-Premuzic, 2011; Gottfried et al., 2016; Shah et al., 2018). Recent advances in neuroscience shed light on the neural mechanisms of curiosity and its links to learning. Gruber, Gelman and Ranganath (2014) show via functional magnetic resonance imaging that the brain's reward system is evoked when people are curious about a phenomenon. This facilitates more enjoyable learning and knowledge retention (deep learning) through memory consolidation.[4] More-

[2]For example, tailoring the level of teaching to children's ability has been shown to be effective in helping those who lag behind to catch up (Banerjee et al., 2007, 2016; Banerji and Chavan, 2016). In highly deprived settings where teacher competency is low, teaching practices as structured as following a written script have been shown to be effective as well; see Gray-Lobe et al. (2022).

[3]Throughout the text, the word curiosity refers to epistemic curiosity, distinguishing human curiosity from animals'. Loewenstein (1994) (and references therein) describes curiosity as reacting positively to new or mysterious events by showing the urge to explore and understand them. Philosopher William describes curiosity as the "impulse towards better cognition" (James, 1983).

[4]Memory consolidation is a process by which acquired information or experiences are poured into

over, they show that once sparked, curiosity creates deep learning moments and enhances the learning of any topic, not only the topic that sparked curiosity initially. While recognized as a powerful motivator for learning, curiosity has not been studied on a large scale within the context of education policy. Our limited knowledge of how to cultivate such a context-dependent trait and the difficulty of measuring it are obvious reasons for the lack of policy-relevant studies. This paper advances the literature on both of these fronts.[5]

The pedagogical program was implemented as two independent randomized controlled trials in two large southern provinces of Turkey. The first trial, implemented in the 2018-2019 academic year in the province of Mersin, included 50 primary schools. We then re-implemented the program in the neighboring province, Adana, recruiting 84 primary schools. This second study took place in the 2021-2022 academic year.[6] Our combined sample includes 134 primary schools with about 11,000 students and 425 teachers. After collecting detailed baseline data from children and teachers in Fall 2018 (Study 1) and Fall 2021 (Study 2), we randomly assigned 78 schools to treatment (25 in Study 1, 43 in Study 2). Teachers from the selected schools received training in the prescribed pedagogy. They were given the entire academic year to practice the pedagogy in the everyday teaching of the curricular topics, with a greater emphasis on science lessons. We collected our endline data in May 2019 (Study 1) and May 2022 (Study 2) to test the effectiveness of the pedagogy using objective test scores, educational aspirations, and a novel incentivized measure of curiosity. When we implemented the second study in the Fall of 2021, we also collected longer-term data from the first study subjects (about three years after the program implementation in 2018).

Curiosity is challenging to measure due to its context-dependent nature. Psychologists use survey tools to elicit different types of curiosity in adults (Litman and Spielberger, 2003; Collins, Litman and Spielberger, 2004; Litman, Collins and Spielberger, 2005; Kashdan et al., 2020). Behavioral tasks are used for very young children (Jirout and Klahr, 2012). Although self-report questionnaires can effectively measure curiosity in adults, it may be necessary to complement them with task-based measurement tools for schoolchildren. We designed an innovative task-based instrument that draws upon the theoretical framework developed by Loewenstein (1994) and insights from neuroscientific research on curiosity. The core idea of our tool is to elicit children's willingness to pay for topic-specific booklets in an incentivized setting. Following the elicitation of willingness to

long-term memory. It is more likely to happen when stimuli spark curiosity; see Gruber and Ranganath (2019).

[5]Recently, psychologists have shown interest in the relationship between what they refer to as "epistemic emotions" and learning. Epistemic emotions include intellectual courage, astonishment, curiosity, interest, wonder, surprise, the joy of verification, and the satisfaction of knowing. These studies are correlational in nature; See Vogl et al. (2019*b*) and Vogl et al. (2019*a*).

[6]Both trials were registered at the AEA Registry before their respective endline dates. The first trial was registered on March 8, 2019, along with a pre-analysis plan. The second trial was registered on November 30, 2021, referring to the first registry for the PAP.

pay and the distribution of booklets, we revisit all classrooms one week later, unannounced, to measure booklet-specific knowledge retention. The distribution of the booklets on the first visit follows one of two regimes. In classrooms that are randomly assigned to the first regime, children receive their preferred booklets based on a randomly determined market price. In classrooms assigned to the second regime, only a random half of the children within each classroom receive booklets, irrespective of the children's willingness to pay or their choice of booklets. By ensuring that the proportion of children receiving booklets and the composition of topics are balanced across treatment status, the second regime enables us to estimate the treatment effect on knowledge retention and explore treatment effects on information sharing and peer learning within the classroom.

Teacher compliance was high in both trials. Almost all teachers reported practicing the prescribed pedagogy, albeit with differing intensity. As aimed, the program significantly changed the way teachers taught curricular topics. Treated teachers reported practicing a more modern, i.e., more learner-centered and inquiry-based teaching relative to control teachers. They also reported a significant increase in their own curiosity level and embracing a growth mindset. We then find that the program significantly improves children's objective test scores in science with no statistically significant impact on math and verbal scores. The estimated effect size on science test scores is about 0.073 standard deviations in the short term. The positive effect on science test scores persists into middle school years, even after a long school closure due to the Covid-19 pandemic. Treated students score 0.073 standard deviations higher than untreated students in a science test covering the middle school curriculum.

The program significantly increases children's willingness to pay for information (curiosity) by about 0.109 standard deviations, implying 5.7% more tokens forgone to purchase a preferred booklet. The effect of the program on the willingness to pay for science-related booklets (scientific curiosity) is similar in size (0.098 standard deviations) and precision. Treated children give up 0.38 more tokens than untreated children for a science-related booklet on average, implying a 11.7% increase in the willingness to pay for scientific information. The effect of the intervention on knowledge retention is striking. Treated children score about 0.114 standard deviations higher in the unannounced booklet test we conducted one week later. Even more striking is that after about three years, including 1.5 years of school closure due to the recent pandemic, treated students score 0.137 standard deviations higher on the same booklet test than untreated children, indicating a remarkably persistent treatment effect on knowledge retention.

We also provide evidence that the program makes friendship networks more effective information dissemination tools. Treated students who did not receive a booklet and whose preferred booklet was received by someone else in their friendship network scored 0.170 standard deviations higher on the booklet test than untreated students in the same condition. We also show that as the information availability increases within friendship networks, treated students exhibit higher

knowledge retention than untreated students. These results strongly suggest more efficient peer learning technology and information dissemination in treated classrooms where students are more curious and passionate about pursuing and sharing knowledge. The improved peer learning is also consistent with the recent evidence that human curiosity is sensitive to the social environment and stimulated by the curiosity of others (Dubey, Mehta and Lombrozo, 2021). Finally, we show that the intervention significantly raises children's aspirations to go to university and study science. While persistent in size, these effects are less precisely estimated in the long run.

Our results suggest that the program's success likely stems from its ability to unleash children's curiosity by changing teaching practices. We rule out improved teacher ability (curricular content knowledge) as a possible mechanism to explain our results. While enhanced curiosity appears as an important channel, we also show that multiple alternative channels may also be at work. We show that the program also increases children's tolerance for uncertainty and makes them more critical in their thinking process.

Our contribution is threefold. First, we evaluate a pedagogical intervention that targets a crucial component of human cognition, curiosity, that has not been studied on a large scale and in a policy-relevant context before. Second, we introduce a novel approach to measuring curiosity in primary school children. Combining the two, we provide suggestive evidence to support the link between childhood curiosity and deep learning in a natural field setting. We show that once sparked, curiosity leads to enhanced knowledge retention in children. Finally, we reveal the learning externalities generated by human curiosity. We show that a pedagogical approach aimed at nurturing students' curiosity not only enhances individual learning outcomes but also promotes information sharing and peer learning within the classroom. These results hold high policy relevance. They can help us design better pedagogical tools to increase pupil and teacher engagement and the quality of learning worldwide. The results are particularly relevant for the developing world, where learning outcomes have been alarmingly low and have deteriorated even further due to the Covid-19 pandemic (Goldhaber et al., 2022).

Our paper relates to several strands of the economics literature. First, by showing the effectiveness of a particular pedagogical approach, it contributes to the literature that strives to improve learning outcomes in developing countries. This literature establishes that school-based inputs have very little effectiveness when not complemented by correct teaching practices (Glewwe et al., 2004; Kremer, Glewwe and Moulin, 2009; Kremer, Brannen and Glennerster, 2013). Related literature explores whether improving teacher motivation and engagement through various incentives improves learning outcomes and yields mixed results (de Ree et al., 2018). Second, the paper also relates to a growing literature that shows that social and emotional skills are likely malleable and can be fostered at young ages (Alan and Ertac, 2018; Alan, Boneva and Ertac, 2019; Alan et al., 2021). We advance this literature by showing that an essential component of human cog-

nition can be cultivated in the classroom through a change in teaching practices. By testing a pedagogy that focuses mainly on science teaching, the paper also speaks to the literature that aims to increase the STEM participation of girls (Buser, Peter and Wolter, 2017; Fischer, 2017; Kahn and Ginther, 2017; Carlana and Fort, 2022). Our heterogeneity analysis reveals that the pedagogy we evaluate increases girls' scientific curiosity more than boys'. Finally, by providing new evidence on the effectiveness of a professional development program, we complement the growing literature on teacher training programs (Popova et al., 2022).

The rest of the paper is organized as follows. Section I summarizes the key features of the program and the context in which it was implemented. Section II details the evaluation design. Section III gives a detailed account of our outcome measures, including our task-based curiosity measure. Section IV describes the data and presents our main results. In Section V, we explore mechanisms through which the program improved knowledge retention and achievement outcomes. We conclude in Section VI.

I. Evaluation Context and The Nature of The Pedagogical Program

The program we evaluate has been developed by an expert team of pedagogy specialists and curricula developers in a private university's innovation center. The program's overarching objective is to promote scientifically informed teaching practices to improve learning outcomes. It aims to do so by replacing traditional teaching with techniques that can stimulate children's curiosity for academic matters. This is especially pertinent in light of the global push for STEM education and better outcomes in science. As such, the program puts a greater focus on the teaching of science.

The Turkish primary school system is designed such that a centrally appointed teacher is assigned to a single classroom in Grade 1 and is expected to teach the same pupils until the end of Grade 4, after which they move on to middle school for Grades 5 to 8 where each subject is taught by a different (branch) teacher.[7] The program has been developed to exclusively benefit primary school teachers, as it is thought that the ideal context for implementing the prescribed pedagogy would be when a single teacher has a full day of contact with their pupils and when science concepts are formally introduced. Such a context is grade 4 of primary school in Turkey.

The intervention was an intensive teacher training program. In training seminars, teachers were first introduced to the concept of deep learning and its connection to epistemic curiosity. The primary objective emphasized throughout the sessions was the importance of tapping into childhood curiosity to enhance children's

[7]While this is the general practice, there are many exceptions to this rule. Firstly, the headteacher can decide which grade level the newly appointed teacher should begin teaching based on the needs of the school. Secondly, the Ministry can re-appoint a teacher, voluntarily or involuntarily, to another school at any grade level. These rotations tend to occur frequently for early career teachers.

learning capacity. Following these intense informational sessions, teachers were introduced to a range of pedagogical practices to foster curiosity in their classrooms. These practices included ways to allow students to suspect and inquire, as well as encourage them to express their interests openly in the classroom. Central to these strategies was capitalizing on children's natural inclination towards mystery, surprise, and humor to capture their attention and create productive teaching moments.

Teachers received a toolkit containing visual and written material to help them practice the pedagogy. These materials are not meant to be a set of materials to be covered in a specified period of time. Rather, they are designed to help the teacher create teachable moments using emotional triggers to hold students' undivided attention before she introduces a new and complex topic. For example, before introducing a science topic on the solar system, which is an official curricular item to be covered, students see a short video on the mysteries of space. The video is designed to capture students' attention, tapping into their love of mystery to create a teachable moment. As another example of creating a teachable moment, this time, using humor, the teacher reads a funny story about a girl who gets excited about exploding liquids before introducing a topic on chemical reactions. While most activities are related to science, the toolkit contains some non-science activities as well. For example, in one of the activities, students read about a fictional student with a deep interest in painting using unconventional tools (finding making a mess with raw eggs liberating). Teachers worked on the toolkit and repeatedly practiced different ways of creating teachable moments during the training seminars with the guidance of education consultants.

We monitored the teachers throughout the implementation process. Every Friday, our designated personnel asked for an update on what was done that week and received pictures of the work done. We sent reminders when we noticed a few weeks of silence in a school. The overall feedback from the teachers regarding the program content was extremely positive. The majority of teachers reported that the program made everyday teaching, not just science teaching, much more enjoyable for children and for themselves. We received reports and visuals from many treated teachers showing their innovative ways of creating teachable moments. Bringing a mysterious box to the classroom that contains valuable information on the layers of the earth, hiding an important piece of information about the phases of matter in the teacher's hair, and hanging the names of the planets in our solar system around an umbrella; are just a few examples.[8] See the Online Appendix C for examples of implementation photographs we received from teachers.

[8]All these topics are part of the 4th-grade Turkish national science curriculum.

II. Evaluation Design

The program was implemented as two independent randomized trials three years apart. The first trial took place in the 2018-2019 academic year, covering 50 primary and 27 post-primary schools in the province of Mersin (Study 1). Due to the logistical difficulties of implementing the program in middle schools, all 27 middle schools were removed from the study at the training sessions.[9] This resulted in a loss of 27 schools, leading to a need for a second trial to enhance the power of the design. We launched the second trial in the 2021-2022 academic year in the neighboring province, Adana, covering 84 primary schools.[10]

In both trials, local authorities provided us with a list of schools in their provinces' socioeconomically deprived neighborhoods. Teachers from these schools were offered participation in the program without any commitment regarding when they would be invited to training seminars. The purpose of this noncommittal invitation was to ensure that we first collect the willingness to participate in the project and then randomize schools into immediate teacher training (treatment group) or training in later academic years (control group). Participation in the program was voluntary on the part of teachers. The program was oversubscribed in both provinces. Due to the large size of Turkish state schools, which generally have multiple classrooms for each grade level, 1 to 6 classrooms were selected randomly from each school for evaluation purposes.[11] Two trials, pooled together, provide us with about 11,000 students and 425 teachers from 134 state primary schools in two large provinces of Turkey. The majority of our sample is composed of 4th graders. We also have some third-grade students in our first study sample.[12]

The timeline of each trial is as follows: We collected baseline data for Study 1 in October 2018, followed by randomization at the school level, stratified by district and grade level. The probability of treatment was 50%, assigning 25 schools to treatment and 25 to control. Teacher training seminars for Study 1 took place in November 2018, and short-term endline data were collected in May 2019. We collected baseline data for Study 2 in October 2021 and conducted the randomization at the school level in the same manner, stratifying by district. We managed to limit our sample to 4th graders in the second study. The ex-ante probability of treatment was again 50%, assigning 43 schools to treatment and 41

[9] With the recommendation of the local authorities in Mersin, we initially included middle schools. The local authority asked us to include 5th-grade students, corresponding to the first year of middle school in Turkey) and their science teachers. However, it became apparent during the training phase that the prescribed pedagogy would be too challenging to implement in a middle school setting.

[10] We first launched the second trial in the 2019-2020 academic year but failed to implement the program and evaluate it due to the school closures caused by the Covid-19 pandemic, which lasted about 1.5 years in Turkey. Therefore, to re-launch the second study, we had to wait until Fall 2021, when the Turkish Ministry of Education opened all schools.

[11] Primary school sizes vary significantly in Turkey, ranging from schools with a single 4th-grade class to overcrowded schools with over 15 classrooms for each grade level.

[12] We admitted a small number of grade 3 classrooms in the first study, comprising about 16% of the sample in this study. This is because we received an overwhelming interest from these teachers and admitted them to the program.

to control in Study 2. Teacher training seminars for 43 treatment schools took place in October 2021. Short-term endline data were collected in May 2022 for this study. The timeline of each study is shown in Figure 1. As can be seen in the figure, at the time we launched the second study in October 2021, we also collected long-term data from our Study 1 subjects in Mersin.

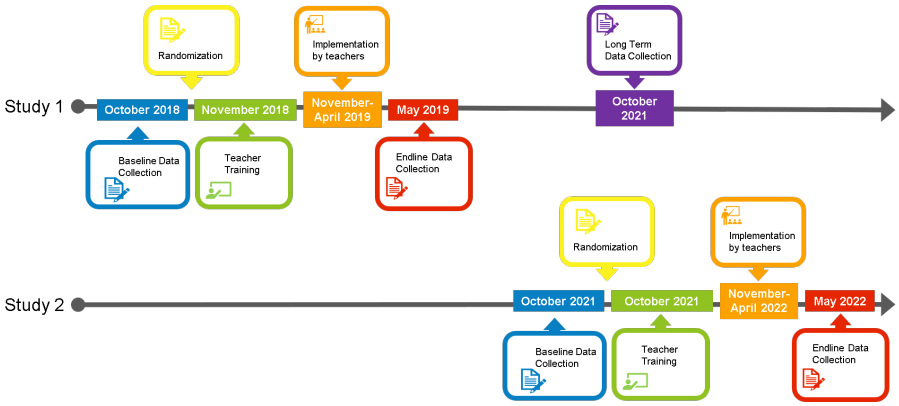


FIGURE 1. TIMELINE OF THE TWO TRIALS

Both baseline and endline data collection were carried out by the research team, assisted by locally recruited and trained field assistants. We made sure that teachers were not present in classrooms during data collection. At baseline, we spent about three lecture hours in each classroom to conduct incentivized games, achievement and psychometric tests, and surveys. We implemented our behavioral curiosity task only at endline. Because of the temporal nature of the task, we organized two visits for each classroom at endline, one week apart. On the first visit, we spent about two lecture hours implementing the curiosity task and collecting other relevant data using tests and surveys. Our second visit was an unannounced surprise visit, which is why our task was implemented only at endline. Upon arrival at the school on the second visit, we kindly asked the teacher to spare us one lecture hour to implement a couple of tests on students and themselves. We will explain the nature of our curiosity task and the tests we implemented later in the text.

In October 2021, almost three years after the first implementation of the program in Mersin (Study 1), we managed to conduct another round of data collection for Study 1. Locating the original subjects of the first study was challenging. While most students were scattered around various middle schools in the same province, some had left the province or left the education system altogether. We eventually located 86% of our original participants with the help of the provincial authority's database. Among those, 84% were formally registered in a state middle school in the province, giving us 72% of our original sample. The attrition is

more likely for girls and refugees, exacerbated by the extended school closures due to the Covid-19 pandemic, but balanced across treatment status (p-value=0.634 and p-value=0.839).[13]

III. Outcomes of Interest

We evaluate the program with respect to a rich set of outcomes using a toolkit comprising achievement tests, surveys, and a novel incentivized task. We first explain our incentivized task.

A. An Incentivized Task to Measure Childhood Curiosity

We designed an incentivized task to capture two prominent aspects of human curiosity: the urge to acquire knowledge and the retention of the acquired knowledge upon satisfying the urge. We benefit from the conceptual framework developed by Loewenstein (1994) for the first component. Based on this framework, we first create a sense of information deprivation in children and then quantify the degree of the urge to acquire information. The second component of our task is informed by the neural mechanisms of curiosity documented in Gruber, Gelman and Ranganath (2014). That is, the higher the urge to know, the stronger the knowledge retention upon satisfying the urge (memory consolidation).

To develop the task, we first conducted extensive pilot surveys and qualitative interviews in several out-of-sample schools to determine the interests of the target age group. Compiling all our survey responses, we identified eight interest categories representing about 95% of all topics of interest. These are, "science", "animals", "history", "human anatomy", "vehicles", "cartoons", "space", and "sports". We then prepared eight small booklets for each topic with a cover that clearly shows the above titles. For example, the cover of the space booklet reads "The mysteries of SPACE," with eye-catching space illustrations to create information deprivation. Figure B2 in the Online Appendix shows the covers of all eight booklets. We placed in each booklet exactly ten pieces of information that are surprising and highly unlikely to be known by children (or by adults). Examples include, "the color of dawn on Mars is blue" in the space booklet, "the actual color of the black box in planes is orange" in the vehicles booklet, or "the shortest battle in history took 38 minutes" in the history booklet.

The implementation of the task in a classroom follows the following steps: We arrive at the classroom with booklets and a basket full of small gift items. The latter are small stationery items that are of value to children of the socioeconomic group we target in this study. We present the booklets to the children one by one, showing the title cover. We tell them that each booklet contains some incredible

[13]Both provinces have a significant refugee population, and all refugee children are covered under the MoE-EU refugee school placement program. However, Turkey's refugee population is highly mobile and difficult to track as they tend to be agricultural laborers. We provide a detailed attrition pattern for Study 1 in Figure B1 in the Online Appendix. A notable number in this figure is 520 missing children the Ministry lost track of in the pandemic period.

facts that are unknown to most people. This step aims to create information deprivation (a strong urge to know) in children. We then ask children to rank these booklets according to their interest in the topic, 1 being the most interesting and 8 being the least interesting.

After obtaining their ranking, we inform children that everyone has an endowment of 10 tokens, and each token can be converted into a gift from our gift basket. We show children these gift items one by one. We then tell them they can also use their tokens to purchase a booklet if they want to. For this, they first need to state the booklet they would like to purchase by ticking the relevant box. We emphasize that they do not have to buy a booklet if they do not want to. In practice, children see 9 options on their screen, 8 topics, and an option of "I don't want a booklet". Then, we begin explaining how this purchase will be made in practice. We first emphasize that all booklets have the same price, and each student can only buy one booklet. We tell them that no one knows the price of a booklet yet, but they need to state their willingness to pay for their preferred booklet, using the options ranging from zero to 10. Then we explain to children that one of two things will happen in their classroom by random chance:

- Market price implementation regime: In this regime, we randomly choose a booklet price (between 1 and 10) for the classroom. Students whose willingness to pay falls under the revealed market price do not receive their desired booklet. They, therefore, convert all their tokens into gift items. Those whose willingness to pay is at or above the revealed market price receive their desired booklets at the market price and convert their remaining tokens into gift items.

- Half-half implementation regime: In this regime, we do not choose a market price for the classroom. Instead, a random half of the classroom receives booklets and all 10 tokens worth of gift items, regardless of their stated willingness to pay and the type of booklet they prefer. The other half of the classroom receives 10 tokens worth of gift items but no booklet. We explain the rationale behind this implementation regime below.

After providing this information and ensuring they fully understand the task, we ask children to state their willingness to pay for their desired booklet with utmost secrecy by tapping the relevant box on their tablet. The elicited willingness to pay, ranging from zero to 10, is our measure of "the urge to know," i.e., curiosity.[14] We conjecture that the treatment will increase children's willingness to pay for information on their preferred topic. Given the program's heavy focus on science, we expect this effect to be particularly prominent in the willingness to pay for science-related booklets, which we refer to as "scientific curiosity." These booklets are science, space, human body, animals, and vehicles.

[14]Willingness to pay elicitation is a standard method in economic research. In the context of information as a good, Hjort et al. (2021) uses this method to elicit policy-makers' willingness to pay for evidence in Brazil.

The novelty of our task lies in its temporal component. In addition to measuring the urge to acquire knowledge, we measure actual knowledge retention using the temporal component of our task. For this, we re-visit all classrooms, unannounced, precisely one week later. In this surprise visit, we give children a 40-question multiple-choice test containing 5 questions from each booklet.[15] The score from this test is our measure of knowledge retention.

Distributing booklets based on children's willingness to pay would generate two confounds for estimating the effect of the treatment on knowledge retention. First, if treatment increases willingness to pay for a booklet, treated classrooms would have more booklets circulating, i.e., have more information available to absorb. Second, if treatment increases interest in science, more science-related booklets would be circulating in treated classrooms, i.e., there would be more science information to absorb. In classrooms assigned to the second regime, only half the children in a given classroom received booklets. In these classrooms, the booklet distribution was random, regardless of children's willingness to pay and their choice of booklets. Ensuring that the proportion of students receiving booklets and the composition of topics are balanced across treatment status, this regime (and only the second regime) allows us to estimate the treatment effect on knowledge retention. This regime also allows us to show the extent to which treatment improves information sharing and peer learning within the classroom.

Allocating a non-zero number of classrooms to the first regime is required to ensure the incentive compatibility of the task. In Study 1, a given classroom had a 50% chance of being subject to either regime, and children were informed accordingly. Because the causal effect of the treatment on information retention can be estimated only in the half-half regime, to improve the power of the experimental design, we implemented the half-half regime in most classrooms (95%) in the second study, and children were informed accordingly. The willingness to pay for a booklet is theoretically independent of the implementation regime, and our data corroborates this: Mean willingness to pay across regimes is statistically not different from each other (p-value=0.484). When implementing this regime, we made sure that every classroom had all 8 booklets. We randomize the regimes within schools to make sure that when we restrict our sample to the classrooms that are subject to the half-half regime, we keep the number of schools (clusters) intact. The Online Appendix D gives full instructions for the task and its implementation.[16]

B. Learning Outcomes and Educational Aspirations

If the program successfully stimulates students' curiosity, we expect deeper learning of curricular topics as well. In particular, given the program's heavy

[15] To do this, we arrived back at schools and kindly asked their permission to take one lecture hour immediately. We gave the same test to teachers and asked them to do their own tests in a quiet, designated room. All our teachers cooperated.

[16] Full implementation kits are available upon request.

emphasis on science teaching, we expect treated students to achieve higher test scores in science. To assess the impact of the program on actual learning outcomes, we implemented tests on math, Turkish (in visit 1), and science (in visit 2) in all classrooms. Because there is no standardized testing system in Turkey for the grade levels we work with, we designed a testing inventory based on the national curriculum.[17] All tests were implemented in classrooms in the absence of teachers both baseline and endline.

In addition to learning outcomes, we assess whether the program affected children's educational aspirations and their plans for study majors. For this, we asked children whether they would like to go to university, and if so, what their aspired topic of study would be. We collected this information both at baseline and endline. We acknowledge that this is not a reliable measure of major choice considering the age of our subjects. Nevertheless, we believe that it gives us an indication of the program's success in raising educational aspirations in children.

IV. Data and Results

We collected data on various cognitive socioemotional skills, beliefs, and preferences at baseline and endline. For students, all demographic information and fluid IQ (Raven and Court, 1998) were measured only at baseline. We conducted standardized achievement tests and elicited risk and ambiguity attitudes using Gneezy and Potters (1997) risky investment task, both at baseline and endline. We collected information via item response surveys to construct measures of epistemic and scientific curiosity (Kashdan and Silvia, 2009), grit (Duckworth and Quinn, 2009), impulsivity (Sleddens et al., 2013), and critical thinking (Sosu, 2013) both at baseline and endline, though critical thinking is measured only in Study 2. Finally, we collected friendship networks both baseline and endline. The motivation to collect these attributes is to establish the validity of our task-based curiosity measure and explore potential channels through which the program might impact learning outcomes. We implemented our curiosity task only at endline.

Our long-term testing inventory was shorter than our short-term inventory because of the constraints imposed by the middle school schedules. We first gathered our students in designated classrooms in their middle schools. Then we gave them the same 40-question booklet test to assess the persistence of our knowledge retention results, followed by math, science, and verbal tests. The last three tests were prepared based on the appropriate grade level covering the national curricula. Finally, we conducted a short survey that elicited curiosity, grit, and aspirations. Table B1 in the Online Appendix shows the variables we collected in each trial at baseline, endline, and long-term follow-up (Study 1 only).

We also collected rich information from teachers. In addition to demographic information, we measured their fluid IQ via Raven's test and their emotional intelligence through the Reading the Mind in the Eyes test (Baron-Cohen et al.,

[17] We benefited from the Ministry's question bank in preparing these questions. We extensively piloted the tests to ensure the appropriateness of the difficulty level.

1997) at baseline. We also collected detailed information regarding teachers' everyday teaching practices and beliefs both at baseline and endline. To measure teacher practices, we adapted some of the item questions from the Teaching and Learning International Survey (TALIS) questionnaire (OECD, 2013) and constructed the following styles: Modern (learner-centered) vs. traditional (lecture-based) teaching, extrinsic vs. intrinsic motivation style, and warm vs. distant (discipline-based) style. For beliefs, we elicited growth mindset (Dweck, 2008), attachment to the profession, competence beliefs, and gender stereotyping. We also measured teachers' curiosity using Kashdan and Silvia (2009) and critical thinking using Sosu (2013). Again the latter was collected only in Study 2. Finally, we tested teachers' curricular knowledge in science to establish whether the intervention increased their content knowledge. We conducted this test in the second (surprise) visit along with the 40-question booklet test.[18] Table B1 in the Online Appendix also gives the variables we collected from teachers. Measurement inventories for students and teachers are presented in the Online Appendix E.

Table A1 presents the balance of student, teacher, and classroom characteristics at baseline. Balance for each study separately is presented in Table B2 and B3 in the Online Appendix. We detect no significant imbalance in any of the variables in either study and conclude that randomization was successful.

We estimate the average treatment effects of the program on outcomes of interest by conditioning on baseline covariates and randomization strata fixed effects:

$$y_{ics} = \alpha_0 + \alpha_1 T_s + X_{ics}^{'}\beta + W_{cs}^{'}\gamma + \delta_d + \varepsilon_{ics} \tag{1}$$

where y_{ics} is the outcome of interest for child i in classroom c, school s. T_s is the binary treatment indicator, which equals one if school s is in the treatment group and zero otherwise, and X'_{ics} is a vector of student-level observables, W'_{cs} is a vector of classroom and teacher level observables measured at baseline. δ_b represents district fixed effects. We chose our covariates by post-double-selection LASSO separately for the short and long term. We defined grade and district dummy variables as partialled-out covariates so that they were not penalized by the LASSO. We also kept gender dummy and fluid IQ scores in the covariate set as we conducted heterogeneity analysis with these variables (specified in our PAP). The short-term covariate set includes gender, fluid IQ, baseline curiosity (survey), refugee status, math and verbal scores, class size, the share of refugees, teacher experience, and the number of children the teacher has. The long-term covariate set is similar to the short-term set but excludes class size and the share of refugees. We also present our main results without covariates; see Table B4, B5, and B6 in the Online Appendix.

The estimated $\hat{\alpha}_1$ is the average intent to treat effect (ITT). Standard errors are clustered at the school level. We also provide wild clustered bootstrapped p-

[18] Both science and booklet tests for teachers were implemented in the second study only.

values in our tables. Throughout the text, we present the results using the pooled sample. The summary of the results for each study separately is given in Figure B3, and detailed results are provided in Tables B7 to B8 in the Online Appendix. We present our results corrected for multiple hypotheses testing (sharpened q-values and Romano Wolf p-values) in Table A2. Most of our results survive the adjustments. We use inverse probability weights for the long-term results (Study 1) to account for attrition.

A. Treatment Effect on Teaching Practices

All treated teachers were expected to practice the proposed pedagogy upon receiving training. Recall that participation in the program was voluntary, and the program was oversubscribed. However, we acknowledge that compliance may not be perfect. To assess compliance, we asked treated teachers to report their estimated degree of program implementation at endline. Specifically, we asked them to mark their estimated degree of implementation using an unmarked 10cm line. The elicited distance gives us a continuous measure of program implementation intensity ranging anywhere between zero and 100%. Note that because this is a pedagogical intervention that aims to influence the way teachers teach, the reported implementation intensity is purely subjective. Nevertheless, we believe that it gives us an idea of teacher compliance. Figure A1 depicts the distribution of the reported implementation intensity for the pooled sample. Overall, treated teachers reported to have accomplished 81% program coverage. Only 4 out of 226 teachers reported zero implementation.

The next question is whether the program changed teachers' classroom practices as intended. Figure 2 plots the estimated treatment effects on teaching styles and beliefs. What emerges from the figure is that the program positively impacted the teaching styles, teachers' epistemic curiosity, and mindset. Treated teachers shifted their teaching practices from traditional lecture-based activities to more modern, learner-centered ones. They also reported less discipline-oriented, warmer interaction with their students. What is remarkable is that teachers themselves became more curious and adopted a more growth mindset. We estimate 0.199 standard deviations higher curiosity and 0.278 standard deviations higher growth mindset for treated teachers than untreated teachers, and both differences are statistically significant at the 1% level.

B. Treatment Effect on Test Scores

Given the effects on teachers' beliefs and classroom practices, we first explore whether the program improved core academic outcomes. Table 1 presents the treatment effects on math, verbal (Turkish), and science test performance. While we do not estimate statistically significant effects on math and Turkish, we find that treated students perform significantly better than untreated students in the science test. The effect size is about 0.073 standard deviations and significant at

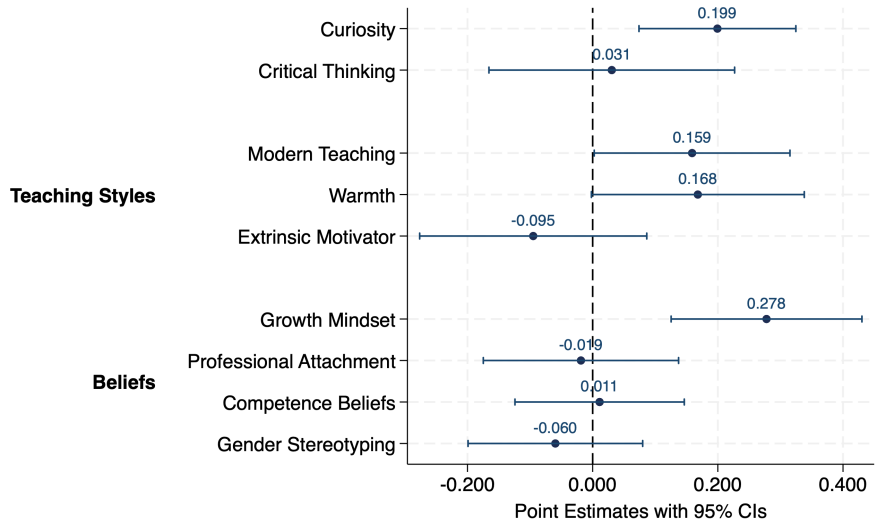


FIGURE 2. TREATMENT EFFECTS ON TEACHERS' PEDAGOGICAL BELIEFS AND TEACHING PRACTICES

Note: The figure depicts the estimated treatment effects on teachers' beliefs, attitudes, and teaching styles. Covariates, selected via post-double-selection LASSO, include baseline beliefs and teaching styles. Grade and district fixed effects included. Standard errors are clustered at the school level.

the 5% level (wild bootstrapped p-value=0.033). The positive effect on science test scores also persists into middle school years. Even three years after the program's implementation, treated students continue to outperform untreated students in science, with an effect size of 0.073 standard deviations, which is significant at the 10% level (wild bootstrapped p-value=0.124). Note that the precision of the long-term estimates is lower due to the smaller sample size.

The near-zero effect sizes observed for math and verbal scores, coupled with the significant and consistent effect on science, may be attributed to the program's strong emphasis on science. This emphasis stems from the fact that science lends itself more readily to this pedagogical approach. Many children find science intriguing and enjoyable, as science topics can be animated and infused with mystery and humor. On the other hand, achieving the same level of engagement in subjects like mathematics and language requires a greater degree of creativity.

The program was also highly cost-effective. Considering printing costs of about 14,000 USD, distribution costs of 3,500 USD, and teacher training costs of about 6,500 USD, the cost per pupil stands at 3.47 USD. Kremer, Brannen and Glennerster (2013) is a valuable reference to put our effect sizes and the program's cost-effectiveness into perspective. The study compares 30 educational RCTs evaluated with respect to learning outcomes. Some of these RCTs are about infrastructure building in highly deprived settings, and they led to significant learning gains. Our intervention is comparable to a smaller set of interventions

Table 1—Treatment Effect on Test Scores

	Short Term			Long Term		
	Science	Math	Verbal	Science	Math	Verbal
Treatment	0.073	0.013	0.032	0.073	-0.017	-0.006
	(0.030)	(0.028)	(0.027)	(0.043)	(0.041)	(0.048)
Wild Bootstrap P-Value	0.033	0.644	0.246	0.124	0.688	0.921
Control Mean	-0.000	-0.000	0.000	-0.000	-0.000	-0.000
Observations	9977	10433	10713	2424	2424	2424
Number of Schools	134	134	134	50	50	50

Note: Estimates are obtained via OLS. The dependent variables are standardized subject test scores. The first 3 columns give short-term results using the pooled sample, and the last 3 provide the long-term results of Study 1. Covariates for the short-term specification, selected via post-double-selection LASSO, include gender, fluid IQ, survey measure of curiosity, refugee status, math and verbal scores as individual baseline characteristics, class size, the share of refugees, teacher experience, and the number of children the teacher has. The long-term covariate set, selected via post-double-selection LASSO, is similar but excludes class size and refugee share. Grade and district fixed effects included. Standard errors are clustered at the school level and are reported in parentheses.

aiming at pedagogy and teacher training.

Pedagogical interventions, such as student tracking, teacher training, and monitoring, offer significant advantages as low-cost policy actions compared to infrastructure and governance interventions. Notably, renowned programs like the Balsakhi program achieved an effect size of 0.14 standard deviations, with a cost-effectiveness ratio of 3.01 standard deviations per $100 (Banerjee et al., 2007). Duflo, Dupas and Kremer (2011) demonstrated an effect size of 0.18 standard deviations and a cost-effectiveness ratio of 34.78 additional standard deviations per $100. The Read-a-Thon Philippines study yielded effect sizes of 0.13 standard deviations immediately after the program and 0.06 standard deviations after three months, displaying a cost-effectiveness ratio of 1.18 standard deviations per $100 (Abeberese, Kumler and Linden, 2014). A recent teacher coaching program in Peru yielded effect sizes ranging from 0.21 to 0.26 standard deviations while giving a cost-effective ratio of 0.04 standard deviations per $100 (Majerowicz and Montero, 2022). Our effect size of 0.075 standard deviations may appear modest compared to these studies. However, our program demonstrates a cost-effectiveness ratio of 2.10 standard deviations per $100, exceeding several other interventions. Additionally, it achieves notable success in generating strong teacher support. In light of these factors, our program stands out as a promising pedagogical intervention.

C. Treatment Effect on the Willingness to Pay and Knowledge Retention

We now investigate whether the program stimulated children's urge to know and their ability to retain knowledge as intended. Before presenting the estimated treatment effects on decisions made in our curiosity task, we establish its

predictive validity. For this, we are interested in two types of predictive validity: First, we want to assess whether the willingness to pay (WTP) for a preferred booklet predicts subsequent knowledge retention. Second, and more generally, whether it predicts core academic outcomes. For the latter, our main interest is the association between willingness to pay for science-related booklets and test scores.

The Predictive Validity of the Willingness to Pay

Table 2—Predictive Power of the Willingness to Pay

Panel A: Raw associations

	Science	Math	Verbal	Retention
WTP (All)	0.044	0.027	0.035	0.048
	(0.019)	(0.017)	(0.020)	(0.015)
WTP (Science)	0.085	0.024	0.086	0.084
	(0.018)	(0.013)	(0.015)	(0.017)
WTP (Non-Science)	-0.048	-0.002	-0.056	0.087
	(0.016)	(0.012)	(0.015)	(0.015)
Observations	4558	4558	4675	4558

Panel B: Raw associations controlling for IQ Score

	Science	Math	Verbal	Retention
WTP (All)	0.023	0.014	0.011	0.040
	(0.016)	(0.015)	(0.016)	(0.015)
WTP (Science)	0.065	0.012	0.061	0.076
	(0.016)	(0.012)	(0.013)	(0.017)
WTP (Non-Science)	-0.046	0.000	-0.052	0.088
	(0.016)	(0.011)	(0.014)	(0.015)
Observations	4558	4558	4675	4558

Note: The table presents OLS coefficients from the regression of the willingness to pay for a booklet separately on test scores (science, verbal, math and booklet test). The analysis uses only the control sample. Standard errors are clustered at the classroom level and are reported in parentheses.

Figure A2 depicts the distribution of forgone tokens for the control sample. Children, on average, forgone 6.14 tokens to receive their desired booklet, with the minimum WTP being zero (7.3% of the control sample) and a maximum of 10 (22.7% of the control sample). Using only the control group, Table 2 presents the predictive power of overall WTP, WTP for science-related booklets, and WTP for non-science related booklet on science, math, and verbal test scores, as well as knowledge retention (performance on the respective questions in the booklet test). Panel A presents raw associations, and Panel B presents the associations controlling for fluid IQ. The results in this table confirm that our WTP measure has reasonable validity in predicting academic outcomes, and it is highly correlated with knowledge retention. Correlations are particularly strong for the willingness

to pay for a science-related booklet (scientific curiosity). A one standard deviation increase in the willingness to pay for a science-related booklet is associated with 0.085 standard deviations higher science, 0.086 standard deviations higher verbal, and 0.024 standard deviations higher math scores, with the first two statistically significant at the 1%, the last at the 10% level. Note that the willingness to pay for either history, sports, or cartoons booklet (non-science curiosity) is negatively associated with academic outcomes but still positively associated with higher knowledge retention (performance in non-science booklet questions). We provide binned scatter plots to show the relationship between WTP and academic outcomes in visual clarity; see Figures B4 to B9 in the Online Appendix.

Table 3—Associations Between Willingness to Pay and Socio-emotional Skills

Panel A: Raw associations

	Curiosity Survey	Science Curiosity Survey	Grit	Impulsivity	Risk	Ambiguity	Critical Thinking
WTP (All)	0.038	0.046	0.051	-0.014	0.228	0.189	0.049
	(0.015)	(0.018)	(0.018)	(0.015)	(0.017)	(0.018)	(0.018)
WTP (Science)	0.048	0.058	0.049	-0.056	0.086	0.075	0.067
	(0.015)	(0.016)	(0.016)	(0.015)	(0.015)	(0.015)	(0.016)
WTP (Non-Science)	-0.015	-0.019	-0.006	0.044	0.108	0.086	-0.027
	(0.014)	(0.016)	(0.016)	(0.014)	(0.016)	(0.015)	(0.016)
Observations	4954	4953	4524	4650	5070	5066	3635

Panel B: Raw associations controlling for IQ Score

	Curiosity Survey	Science Curiosity Survey	Grit	Impulsivity	Risk	Ambiguity	Critical Thinking
WTP (All)	0.029	0.035	0.044	-0.006	0.231	0.193	0.040
	(0.015)	(0.017)	(0.018)	(0.015)	(0.017)	(0.018)	(0.018)
WTP (Science)	0.038	0.047	0.041	-0.048	0.090	0.078	0.058
	(0.015)	(0.015)	(0.016)	(0.015)	(0.015)	(0.015)	(0.016)
WTP (Non-Science)	-0.014	-0.017	-0.004	0.042	0.107	0.086	-0.025
	(0.014)	(0.016)	(0.015)	(0.014)	(0.016)	(0.015)	(0.016)
Observations	4954	4953	4524	4650	5070	5066	3635

Note: The table presents OLS coefficients from the regression of the willingness to pay for a booklet separately on curiosity, scientific curiosity, grit, impulsivity, risk and ambiguity tolerance, and critical thinking. Risk and ambiguity tolerance is measured via incentivized tasks. Other skills are measured via item-response questionnaires. All measures are standardized. The analysis uses only the control sample. Standard errors are clustered at the classroom level and are reported in parentheses.

Table 3 further validates our incentivized task. Here, we check whether the willingness to pay for a booklet correlates with survey measures of curiosity developed by Kashdan and Silvia (2009). In addition, we conjecture that curiosity may be correlated with attitudes toward uncertainty, grit, critical thinking, and impulsive behavior, acknowledging its possible relationship with other social and emotional skills we do not measure in this paper. Panel A presents raw associations, and Panel B presents the associations controlling for fluid IQ. We observe

strong positive correlations between our curiosity and established survey measures of curiosity. Moreover, the willingness to pay for science-related booklets (our scientific curiosity measure) correlates positively with grit, critical thinking, and risk and ambiguity tolerance and negatively correlates with impulsivity.

Table B9 in the Online Appendix shows the associations between the WTP and academic outcomes, controlling for all cognitive and socioemotional skills available in our data. The table shows that fluid IQ is the most significant predictor of academic success. A one standard deviation increase in fluid IQ is associated with a 0.414 standard deviations gain in science test scores. Over an above IQ, we observe significant predictive power coming from grit, impulsivity, and critical thinking in the expected direction. More importantly for the validity of our measure, we observe that the willingness to pay for a science-related booklet is still highly predictive of science and verbal test scores, even after controlling for IQ and all other socioemotional skills available in our data, with the estimated size of 0.068 standard deviation, significant at the 1% level. We also provide the associations between the WTP and socio-demographic characteristics in Table B10 in the Online Appendix.

Treatment Effect on Interest and the Willingness to Pay for a Booklet

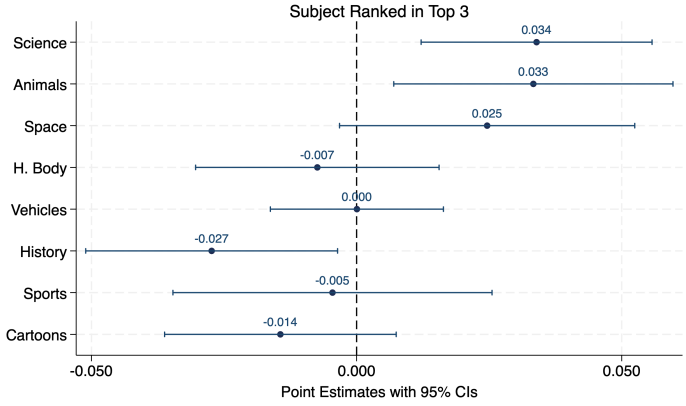


Figure 3. Treatment Effect on the Ranking of Topics

Note: The figure depicts the average marginal treatment effects obtained from logistic regressions on subject ranking. The dependent variables are binary indicators of one if the respective booklet is ranked as one of the top 3 interests by the student. Covariates, selected via post-double-selection LASSO, include gender, fluid IQ, survey measure of curiosity, refugee status, math and verbal scores as individual baseline characteristics, class size, the share of refugees, teacher experience, and the number of children the teacher has. Grade and district fixed effects included. Standard errors are clustered at the school level.

Recall that students' first decision in the curiosity task was to rank the presented eight topics from the most interesting to the least. Figure 3 depicts the average

marginal treatment effects on the probability of a given topic being ranked as top 3. Treated children are 3.3 percentage points more likely to rank science and animals as their top 3 interests, with the former statistically significant at the 1% and the latter at the 5% level. Interest in space is also higher among treated children than control, but the effect is less precisely estimated (p-value=0.083).

TABLE 4—TREATMENT EFFECT ON THE CHOICE OF BOOKLET AND WILLINGNESS TO PAY

Panel A: Choice of Booklet

	Science Related	Non-Science Related	No booklet
Treatment	0.038	-0.009	-0.029
	(0.011)	(0.011)	(0.008)
Wild Bootstrap P-Value	0.000	0.406	0.001
Control Mean	0.495	0.440	0.065
Observations	10898	10898	10898
Number of Schools	134	134	134

Panel B: Willingness to Pay

	WTP (All)	WTP (Science)	WTP (Non-Science)
Treatment	0.109	0.098	-0.009
	(0.040)	(0.026)	(0.025)
Wild Bootstrap P-Value	0.017	0.001	0.708
Control Mean	-0.000	0.000	-0.000
Observations	10892	10891	10891
Number of Schools	134	134	134

Note: Estimates are obtained via OLS. Panel A reports the estimated effects on the choice of a booklet. The dependent variables are binary indicators of choosing a science-related booklet (science, space, vehicles, human body, and animals) in column 1, choosing a nonscience-related booklet (history, sports, and cartoons) in column 2, and choosing no booklet option in column 3. Panel B reports estimated effects on the WTP for a booklet, WTP for a science-related booklet, and WTP for a non-science booklet. Covariates, selected via post-double-selection LASSO, include gender, fluid IQ, survey measure of curiosity, refugee status, math and verbal scores as individual baseline characteristics, class size, the share of refugees, teacher experience, and the number of children the teacher has. Grade and district fixed effects included. Standard errors are clustered at the school level and are reported in parentheses.

Table 4 Panel A shows the program's impact on topic interest further by focusing on children's preferred booklet, i.e., the booklet for which they stated their willingness to pay. The first column shows the treatment effect on the probability of choosing to purchase a science-related booklet (science, animals, space, vehicles, human anatomy). The second column presents the treatment effect on choosing a non-science booklet (history, sports, and cartoons). The last column gives the estimated effect of the treatment on "no interest," i.e., the probability of choosing not to purchase a booklet. Notice that about 50% of the children in the control group stated their willingness to purchase a science-related booklet. This value goes up by about 4 percentage points in the treatment group, and

this difference is statistically significant at the 1% level.[19] It appears that the program shifted children's interest to science topics but not much at the expense of non-science topics (see column 2). As shown in column 3 of the table, the program lowered the probability of no interest, i.e., stating zero willingness to pay, by 2.9 percentage points, representing about a remarkable 45% effect.

Table 4 Panel B presents the estimated treatment effects on the willingness to pay for the desired booklet. We standardize WTP to have a mean zero for the control group, so the coefficient estimates are standard deviation effects. Column 1 presents the overall willingness to pay for any preferred booklet, column 2 for a science-related booklet, and the last column for a non-science booklet. Note that the measure sets the willingness to pay for unpreferred booklets to zero. This could potentially pose a threat to internal validity for WTP (science) and WTP (non-science) if the program, rather than enhancing scientific curiosity, shifts the interest of already curious children from non-science to science topics. However, we do not see evidence of such substitution in Panel B.

We estimate a significant 0.109 standard deviation effect on overall willingness to pay. In terms of tokens, this corresponds to the willingness to forgo about 0.35 extra tokens for the preferred booklet. Given that children forgo 6.1 of their tokens on average for their preferred booklets in the control group, this effect implies a 6% treatment effect and is significant at the 5% level based on the wild bootstrapped p-value (0.017). The effect on the willingness to pay for a science-related booklet (scientific curiosity) is similar with about 0.098 standard deviation treatment effect, again precisely estimated (wild bootstrapped p-value of 0.001). The estimated effect on the willingness to pay for non-science booklets is statistically zero. These results indicate that the program is successful in stimulating children's interest in science-related topics and their curiosity in general. Our next question is whether this stimulated urge to know translates into actual learning. The temporal component of our task and the half-half implementation of booklet distribution allow us to answer this question.

Treatment Effect on Knowledge Retention

The estimated treatment effects on the willingness to pay suggest that in the market price regime, where the price of a booklet is determined randomly, treated classrooms necessarily end up with a proportionally higher number of booklets. This means that treated classrooms have more information (booklets) available for all, making it more likely to acquire and retain the knowledge available in the classroom. A clean identification of the effect of the program on knowledge retention requires the amount and the content of information to be balanced across treatment status. The half-half implementation regime delivers this by design. Recall that in classrooms subject to this regime, we distributed the booklets randomly to half of the students regardless of their willingness to pay and their choice

[19] 53 students stated that they did not want a booklet but still stated their WTP, indicating they did not fully understand the task. We do not exclude these students from our sample.

of booklets. Therefore, we can compare the performance on the surprised booklet test across treatment status by restricting our sample to the classrooms that were subject to the half-half regime. Panel A in Table 5 presents the estimated treatment effects on booklet test scores. The first 3 columns give short-term, and the last 3 give long-term effects (only Study 1).

TABLE 5—TREATMENT EFFECT ON KNOWLEDGE RETENTION

Panel A: Knowledge Retention

	Short Term			Long Term		
	Retention	Science Retention	Non-Science Retention	Retention	Science Retention	Non-Science Retention
Treatment	0.114	0.102	0.086	0.137	0.156	0.056
	(0.051)	(0.046)	(0.044)	(0.062)	(0.052)	(0.064)
Wild Bootstrap P-Value	0.027	0.026	0.069	0.065	0.016	0.407
Control Mean	-0.000	0.000	0.000	-0.000	0.000	0.000
Observations	9070	9070	9070	1335	1335	1335
Number of Schools	134	134	134	50	50	50

Panel B: Knowledge Retention (excluding Preferred Booklet)

	Short Term			Long Term		
	Retention	Science Retention	Non-Science Retention	Retention	Science Retention	Non-Science Retention
Treatment	0.119	0.105	0.097	0.169	0.154	0.120
	(0.053)	(0.047)	(0.044)	(0.061)	(0.053)	(0.069)
Wild Bootstrap P-Value	0.028	0.026	0.040	0.017	0.016	0.130
Control Mean	-0.000	0.000	-0.000	-0.000	-0.000	0.000
Observations	8299	8299	8299	1218	1218	1218
Number of Schools	134	134	134	50	50	50

Note: Estimates are obtained via OLS using the sample restricted to the half-half regime. The dependent variables are standardized booklet test scores (knowledge retention). The first three columns give short-term results using the pooled sample, and the last three provide the long-term results obtained from Study 1. Covariates for the short-term specification, selected via post-double-selection LASSO, include gender, fluid IQ, survey measure of curiosity, refugee status, math and verbal scores as individual baseline characteristics, class size, the share of refugees, teacher experience, and the number of children the teacher has. The long-term covariate set, selected via post-double-selection LASSO, is similar but excludes class size and refugee share. Grade and district fixed effects included. Standard errors are clustered at the school level and are reported in parentheses.

The impact of the program on the ability to retain knowledge is striking. Treated students performed about 0.114 standard deviations higher than untreated students overall, and the performance difference is similar in science topics (0.102 standard deviations). Both effects are significant at the 5% level. Note that treated students performed better even in non-science topics of the test. Moreover, they continued to exhibit higher booklet knowledge after three years, even after disruptions caused by the Covid-19 pandemic. Treated students performed 0.137 standard deviations higher in the booklet test 3 years after the intervention. The retention of science-related topics after 3 years is about 0.156 standard deviations, and this effect is significant at the 5% level (wild bootstrapped p-

value=0.016), whereas the retention effect on non-science topics fades in the long term.

We hypothesize that enhanced curiosity leads to the absorption of the available information in one's environment, whether such information is of interest to the individual or not. Our research design allows us to test this hypothesis. We do know the student's preferred booklet, whether or not she received a booklet, and if received, which booklet she received. We constructed a booklet test score performance for each child by eliminating the questions related to her preferred topic. Panel B presents estimated effects on knowledge of topics outside students' preferred booklet. The overall retention results refer to topics other than the preferred one. Science retention refers to the performance of students who preferred a non-science booklet on science-related topics. Non-science retention refers to the performance of students who preferred a science booklet on non-science-related topics. We see positive and significant effects with similar magnitudes presented in Panel A, both in the short and the long term.

Note, however, that to the extent that treated students have a preferred booklet or tend to prefer science booklets more, the composition of questions we base our testing on will differ across treatment and control. Recall that 6 percent of students in the control group stated that they did not want a booklet. This value is 45% less in the treatment group, meaning more control students are tested on all 8 booklets (40 questions rather than 35 questions) than treatment, giving them the advantage of getting more questions correct, working against finding positive treatment effects. Despite this, these results should be considered suggestive rather than causal. These results also clue us in on a possible social aspect of curiosity and learning, which we explore further in the next section.

Treatment Effect on Information Dissemination in the Classroom

It has been shown that in addition to being associated with deep learning, human curiosity has positive externalities. Hartung and Renner (2013) and Litman and Pezzo (2007) show that curiosity is associated with passionate information sharing. Dubey, Mehta and Lombrozo (2021) show that human curiosity is sensitive to the social environment and stimulated by the curiosity of others. These externalities imply enhanced peer learning in our context, and our research design allows us to test the presence of these externalities. Our test involves exploring whether the program made the classroom a denser learning environment where students share what they learn with their peers. We collected friendship networks at baseline and endline by asking each student to nominate at most three peers in their classrooms as their friends. With these nominations and the fact that we know who received which booklet, we can gain a deeper understanding of how the information provided to a subset of students in the classroom is disseminated and how treatment interacts with the way information is disseminated.

Table 6 shows the treatment effect on retention for students who received a booklet (Panel A) and those who did not (Panel B). The former takes the stu-

TABLE 6—TREATMENT EFFECT ON KNOWLEDGE RETENTION THROUGH INFORMATION DISSEMINATION

Panel A: Booklet Received

	Retention	Science Retention	Non-Science Retention
Treatment	0.150	0.129	0.119
	(0.058)	(0.053)	(0.050)
Wild Bootstrap P-Value	0.007	0.013	0.026
Control Mean	0.000	-0.000	-0.000
Observations	4217	4217	4217
Number of Schools	134	134	134

Panel B: No Booklet Received

	Retention	Science Retention	Non-Science Retention
Treatment	0.080	0.070	0.061
	(0.048)	(0.043)	(0.043)
Wild Bootstrap P-Value	0.125	0.123	0.183
Control Mean	0.000	0.000	0.000
Observations	5283	5283	5283
Number of Schools	134	134	134

Panel C: Network Effect

	Retention	Science Retention	Non-Science Retention
Treatment	0.170	0.145	0.138
	(0.076)	(0.072)	(0.073)
Wild Bootstrap P-Value	0.036	0.057	0.082
Control Mean	0.000	-0.000	0.000
Observations	1054	1054	1054
Number of Schools	134	134	134

Note: Estimates are obtained via OLS. The dependent variables are standardized booklet test scores (knowledge retention). Panel A uses the sample of booklet recipients only in the half-half regime. Panel B uses the sample of students who did not receive a booklet. Panel C uses the sample of students who did not receive a booklet but have at least one person in their network who has received the booklet of their choice. Covariates, selected via post-double-selection LASSO, include gender, fluid IQ, survey measure of curiosity, refugee status, math and verbal scores as individual baseline characteristics, class size, the share of refugees, teacher experience, and the number of children the teacher has. Grade and district fixed effects included. Standard errors are clustered at the school level and are reported in parentheses.

dents who received booklets under the half-half regime, and the latter uses all students who did not receive any booklet. Recall that not having a booklet may be due to a pure chance (under half-half regime), or by unwillingness to buy a booklet (under market price regime), or by falling under the market price (under market price regime).[20] The effect sizes are larger and more precisely estimated for booklet recipients. Treated students who received a booklet performed 0.150 standard deviations better on the booklet test than their untreated counterparts. This effect is significant at the 1% level. However, the retention effects are generally weaker for those who did not receive a booklet. Treated students who did not receive a booklet performed 0.080 standard deviations better in the booklet test than their untreated counterparts, but this effect is statistically weak (wild bootstrapped p-value=0.125).

In Panel C, we present the retention results for students who did not receive a booklet and whose preferred booklet was received by someone else in their friendship network. Here, the friendship network of a student contains all her friendship nominations (out-degree ties) and all her classmates who nominate her as their friend (in-degree ties).[21] The results are remarkable: We estimate 0.170 standard deviations higher booklet knowledge for these students overall, suggesting a significantly higher pursuit of information among treated students. Treatment effects on science and non-science knowledge retention are 0.145 and 0.138 standard deviations for these students, respectively. Note that while highly restricted, this sample is balanced across treatment status with respect to baseline characteristics; see Table B12 in the Online Appendix. Finally, the retention effects presented in Panel A and C are not statistically different.

We also explore the treatment effect heterogeneity under differential information availability within friendship networks to complement these results. Figure 4 plots the estimated treatment effects on knowledge retention conditional on information availability within friendship networks. Here, we focus on the information the student is interested in, i.e., booklets that he/she ranked as top 3. Panel A presents estimated effects conditional on receiving a booklet and an increasing number of top-3 ranked booklets available in the friendship network (zero, one, two, or three and more booklets). Panel B presents estimated effects conditional on receiving no booklet and an increasing number of top-3 ranked booklets available in the friendship network. The depicted treatment effects suggest significantly higher knowledge retention for treated children, monotonically increasing with the availability of information within their networks. Consistent with Table 6 Panel A, the estimated effects are stronger for booklet owners. Treated booklet owners who are the sole booklet owners within their network per-

[20] Here, the underlying hypothesis is that information will flow to those who did not receive a booklet, either by chance or due to their lower willingness to pay. The estimates using only the half-half regime are not materially different; see Table B11 in the Online Appendix.

[21] We checked whether the program had any impact on the network structure, such as the network density, the number of friendship ties, the number of isolated students, and the number of reciprocal ties, and found no such evidence.

form 0.064 standard deviations better in science-related booklet questions than untreated booklet owners in the same situation. This effect is statistically insignificant. The estimated treatment effect goes up to 0.278 standard deviations for this group when their friendship network possesses more than three science-related booklets. While we cannot reject the equality of these effects due to insufficient power, the visible monotonicity is important to note.

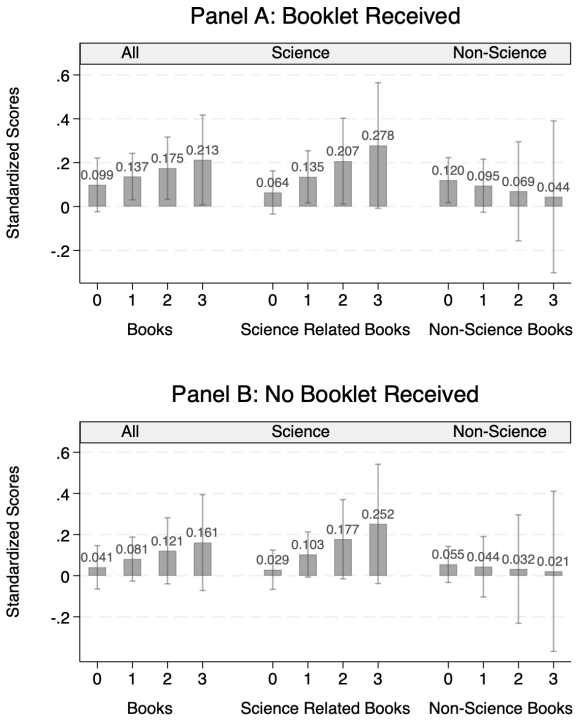


Figure 4. Information Availability and Treatment Effects on Knowledge Retention

Note: The figure depicts the estimated treatment effects on standardized booklet test scores (knowledge retention). Panel A restricts the sample to those who received a booklet, and Panel B to those who did not receive a booklet. Depicted coefficient estimates are obtained by further restricting each sample as having none, one, two, and more than three top-3 ranked booklets in the student's network (our measure of information availability within the friendship network). Covariates, selected via post-double-selection LASSO, include gender, fluid IQ, survey measure of curiosity, refugee status, math and verbal scores as individual baseline characteristics, class size, the share of refugees, teacher experience, and the number of children the teacher has. Grade and district fixed effects included. Standard errors are clustered at the school level.

The effects are weaker for those who did not receive a booklet (Panel B), but we still observe monotonically increasing treatment effects on knowledge retention in science-related topics as information availability increases within the network.

We interpret these estimates as more efficient information dissemination and peer learning in treated classrooms where students are more curious and passionate about pursuing and sharing knowledge. Note also that stronger effects estimated for the booklet owners suggest that access to available information within networks via booklet exchange is more prominent in treated classrooms. Put differently, treated booklet owners, who are in a better position to access other booklets in their network, utilize this access better and absorb more information than their control counterparts.

D. Treatment Effects on Educational Aspirations

Table 7—Treatment Effect on Aspirations

Panel A: Short Term

	University	Science	Engineering	Medical	Non-STEM
Treatment	0.008	0.023	0.000	-0.002	-0.021
	(0.005)	(0.007)	(0.007)	(0.008)	(0.011)
Wild Bootstrap P-Value	0.113	0.003	0.937	0.799	0.061
Control Mean	0.950	0.115	0.118	0.162	0.605
Observations	10721	10212	10212	10212	10212
Number of Schools	134	134	134	134	134

Panel B: Long Term

	University	Science	Engineering	Medical	Non-STEM
Treatment	0.010	0.018	0.011	-0.015	-0.014
	(0.008)	(0.019)	(0.016)	(0.018)	(0.022)
Wild Bootstrap P-Value	0.230	0.385	0.536	0.431	0.593
Control Mean	0.950	0.129	0.118	0.214	0.540
Observations	2318	2181	2181	2181	2181
Number of Schools	50	50	50	50	50

Note: Estimates are obtained via OLS. The dependent variables are binary choice variables of intention to go to university, intention to choose a science major, engineering major, medicine, and non-STEM major. Panel A presents short-term results from the pooled sample, and Panel B long-term results from Study 1. Covariates for the short-term specification, selected via post-double-selection LASSO, include gender, fluid IQ, survey measure of curiosity, refugee status, math and verbal scores as individual baseline characteristics, class size, the share of refugees, teacher experience, and the number of children the teacher has. The long-term covariate set, selected via post-double-selection LASSO, is similar but excludes class size and refugee share. Grade and district fixed effects included. Standard errors are clustered at the school level and are reported in parentheses.

Given the program's focus on science and its positive impact on interest in science, it is plausible that it also affected children's educational aspirations. To measure aspirations, we asked the children two questions. First, we asked whether they intended to go to university when they grew up. Second, if they did, we what study major they wanted to pursue. For the latter, we gave them a full list of study majors to choose from. The first column of Table 7 presents the estimated

treatment effect (average marginal effect) on the willingness to go to university. The following columns present the estimated average marginal effects on planned study majors. These are science, engineering, medicine, and Non-STEM (social sciences and humanities). Note first that almost all (95%) children in the control group stated that they plan to go to university when they grow up. We estimate a statistically weak treatment effect of 1 percentage point on this high base. More importantly, only 11.5% of the children in the control group state their plan to major in science at university. This value is 2.3 percentage points higher for the treatment group, implying a 20% treatment effect. We estimate null effects for engineering and medicine. The estimated negative effect on non-STEM majors suggests that the positive effect we estimate for science comes at the expense of non-STEM majors. While 61% of the students express a preference toward a social science topic in the control group, treated students are 0.02 percentage points less likely to state such a preference. Note, however, while estimated sizes remain similar in the long run, they are estimated imprecisely, likely due to insufficient statistical power.

E. Heterogeneity in Treatment Effects

As stated in our PAP, we explore heterogeneity in treatment effects with respect to two characteristics. First, we check whether the estimated effects are different across gender. Second, we investigate whether the program has a differential impact on children with different levels of cognitive ability (fluid IQ). The first thing to report regarding gender heterogeneity is that we observe more non-science curiosity for boys (mainly driven by their interest in sports) with no gender difference in the willingness to pay for a science-related booklet; see Table B10 in the Online Appendix for the associations between WTP and demographic characteristics in the control group.

Interestingly, as seen in the first panel in Table A3, the program's effect on the shift toward science topics mainly comes from girls. Treated girls are 7.7 percentage points more likely to choose a science-related booklet relative to untreated girls. The corresponding estimate is statistically zero for boys. As for choosing no booklet (no interest), we estimate no gender heterogeneity. Both boys and girls in the treatment group are significantly less likely to choose "no booklet" than those in the control group, suggesting that the program stimulated the overall interest of both boys and girls. Similarly, we detect a significant gender heterogeneity in the treatment effect on the willingness to pay for a booklet. The estimates in Panel B indicate that while the program is effective in increasing curiosity for both genders, the results seem stronger for girls. Treated girls have 0.171 standard deviations higher willingness to pay for a science-related booklet than untreated girls. We reject the equality of effects across gender for overall curiosity as well as science and non-science curiosity. We do not detect any noteworthy gender heterogeneity with respect to knowledge retention, test scores and educational aspirations; see Tables A4 to A6.

Tables B13, B14, B15, and B16 in the Online Appendix present treatment effect heterogeneity with respect to cognitive ability. Here, we use our measure of fluid IQ (Raven score) and estimate treatment effects separately for high (above median) and low IQ (below median) levels. Overall, the estimated effects seem stronger for students with higher cognitive ability, although we fail to reject the equality of the estimated effects in most cases. The exception is the treatment effect on the willingness to pay; see Table B13 in the Online Appendix. As can be seen in Panel B, while the program seems effective in increasing the willingness to pay for both IQ levels, its effect is stronger for students with high IQs. This is reflected in the retention results (Table B14 in the Online Appendix, especially for the long-term results. Treated high IQ children performed 0.216 standard deviations higher than control children in the booklet test given to them by surprise three years after the intervention. We do not estimate any treatment effect heterogeneity aspirations with respect to IQ; see Table B16 in the Online Appendix.

Taken together, our results suggest that the program was highly successful in increasing children's interest in science and stimulating their curiosity. In addition, it was highly effective in enhancing children's ability to retain the acquired knowledge and improving science test scores. The next section will explore possible mechanisms through which the program achieves these positive results.

V. Potential Mechanisms

While the program had a specific focus on stimulating curiosity, given the correlations, we established in Table 3, it is plausible that it also affected related attributes in children, potentially leading to improved learning. To investigate this, we explore whether the program had any impact on other attributes correlated with curiosity.

Figure 5 depicts the estimated treatment effects on grit, impulsivity, risk and ambiguity tolerance, critical thinking, and survey measure of epistemic and scientific curiosity. For the long term (Study 1), we only have self-reported epistemic and scientific curiosity and grit. Note first that consistent with the effects we estimate on the behavioral task. We estimate a 0.207 standard deviation treatment effect on self-reported curiosity and a 0.161 standard deviation effect on self-reported scientific curiosity. The former effect persists into adolescence, but the latter does not (Study 1). We also find that treated children have become more tolerant of risk and ambiguity and more critical in their thinking process than untreated children. We do not estimate a statistically significant treatment effect on grit in the short term but observe an effect of 0.070 standard deviation on grit in treated children in the long term. The latter effect is significant at the 10% level.

These results suggest that while it is plausible that the enhanced curiosity explains our knowledge retention results, other mechanisms may be at work for the improved science test scores. For example, in addition to enhanced curiosity, the program's positive impact on children's critical thinking skills may be

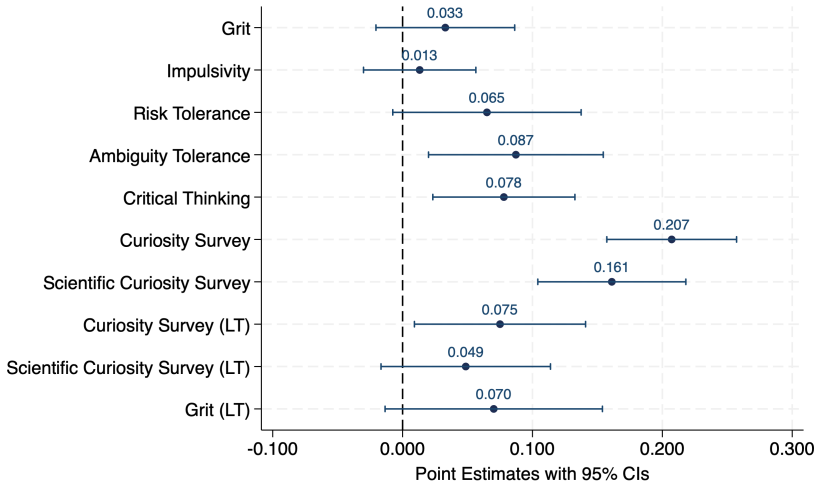


FIGURE 5. TREATMENT EFFECTS ON STUDENTS' BELIEFS AND ATTITUDES

Note: The figure depicts the estimated treatment effects on children's socio-emotional skills, beliefs, and attitudes. Covariates for the short-term specification, selected via post-double-selection LASSO, include gender, fluid IQ, survey measure of curiosity, refugee status, math and verbal scores as individual baseline characteristics, class size, the share of refugees, teacher experience, and the number of children the teacher has. The long-term covariate set, selected via post-double-selection LASSO, is similar but excludes class size and refugee share. Grade and district fixed effects included. Standard errors are clustered at the school level.

responsible for improved test scores. Recall that the program had a significant impact on teachers' classroom practices, swaying teachers toward adopting a more inquiry-based, learner-centered teaching style. While this style of teaching may have stimulated children's curiosity about science-related topics, it may also have led them to be more critical in their thinking process, which may be partially responsible for improved test scores.

While we refrain from fully subscribing to a particular channel, we can safely rule out one. An alternative mechanism for the improved science test scores could be that the program improved teachers' curricular content knowledge. Even though the program did not provide any curricular information, treated teachers may have invested some time to increase their ability in science as they became more curious. A similar mechanism may explain the retention results. Upon observing them circulating in the classroom, teachers may have learned the information provided in booklets out of curiosity and taught their students, even though we gave no indication that we would return and give a test containing questions about the information provided in booklets. We estimate precise null effects on teachers' curricular knowledge in science (p-value=0.908), measured by their performance on a test covering the 4th-year science curriculum. We also find

no evidence of higher booklet knowledge in treated teachers (p-value=0.456).[22]

VI. Conclusion

We demonstrate the effectiveness of a pedagogical program aimed at fostering children's curiosity in the classroom. The pedagogy is informed by recent research on the neural mechanisms of human curiosity and mainly targets science teaching in elementary schools. The program offers teachers practices that help them create teachable moments by tapping into children's natural love of learning. The program was implemented as two independent clustered randomized controlled trials in two large provinces of Turkey, involving 134 primary schools, 425 teachers, and about 11,000 children of age 9 to 11. The program made a significant impact on teachers' teaching practices, shifting from traditional lecture-based instruction to learner-centered inquiry-based approaches. Moreover, the pedagogy significantly improved children's objective test scores in science, with the effects persisting well into early adolescence.

To evaluate the program's effect on children's curiosity, we developed a novel behavioral measure that quantifies children's urge to acquire knowledge and their ability to retain knowledge upon satisfying the urge for an extended period. Using this measure, we found that the intervention increases children's curiosity, measured by their willingness to pay for information and their ability to retain knowledge. Furthermore, our design allowed us to show that classroom practices that nurture children's curiosity also encourage more information sharing in the classroom, revealing the link between pedagogy and peer learning.

The results are promising and likely to hold high external validity for two reasons. First, despite the fact that the participation was voluntary, the program was oversubscribed. Most teachers were eager to join the program in almost all contacted schools, implying generalizability within the Turkish context. Second, Turkey is a middle-income (OECD) country that does not face challenges related to infrastructure, teacher absenteeism, and low teacher competency, as we often observe in low-income countries. Turkish teachers are reasonably well-educated and value pedagogical training. Therefore, our results are likely to be generalized to settings such as Southern and Eastern Europe, Latin America, and some relatively well-off Asian countries facing educational challenges similar to the challenges facing Turkey.

Global learning poverty is at its worse in the wake of the devastating Covid-19 pandemic. While the learning crisis predates the pandemic, the pandemic-related school closures made matters disproportionately worse for underprivileged children. They further widened the already sizeable socioeconomic achievement gaps to an alarming level in both developed and developing countries. The crisis now calls for evidence-informed and scalable actions more urgently than ever.

[22] Estimates for science test scores and booklet test scores for teachers are available only for Study 2 as we were not allowed to test teachers in the first study.

One action may be to equip teachers with effective teaching practices that have a high chance of increasing teacher and pupil engagement, resulting in quality learning. We provide rigorous evidence on the effectiveness of one such scalable and cost-effective action. We envision a couple of ways this program can be scaled up. One way is through incorporating the training in regular professional development seminars given to teachers at the beginning of the academic year. Another way can be to offer seminar courses for teacher candidates in universities. It is unclear which delivery medium would be more effective and may be a topic of future research.

Appendix A: Tables and Figures

Table A1—Balance at Baseline

	N	Control Mean	Treatment Mean	Diff pvalue
Student Characteristics				
Male	13039	0.510	0.509	0.961
Age in Months	13039	112.433	112.745	0.254
Fluid IQ Score	10912	-0.017	0.015	0.824
Math Score	10922	-0.017	0.015	0.897
Verbal Score	10922	-0.016	0.014	0.964
Curiosity	13039	-0.008	0.007	0.641
Risk Attitude	13039	2.615	2.580	0.770
Ambiguity Attitude	10409	2.468	2.425	0.845
Gender Roles	10613	0.035	0.027	0.386
Home - Computer	10758	0.501	0.523	0.513
Home - Internet	10738	0.796	0.797	0.721
Siblingship Size	10814	2.737	2.722	0.902
Birth Order	10814	2.611	2.591	0.995
Teacher Characteristics				
Male	425	0.271	0.288	0.678
Age	425	45.487	44.659	0.248
Fluid IQ Score	425	17.759	17.704	0.809
Cognitive Empathy Score	425	23.046	22.944	0.868
Married	425	0.829	0.850	0.619
Number of children	425	1.814	1.712	0.184
Teaching experience in Years	425	21.005	20.504	0.411
University Graduate	425	0.940	0.951	0.504
Curiosity	425	-0.051	0.084	0.143
Gender Styping Beliefs	425	-0.006	-0.011	0.863
Growth Mindset	425	-0.041	-0.002	0.691
Professional Attachment	425	-0.012	-0.025	0.865
Competence Beliefs	424	-0.021	0.055	0.367
Modern Teaching	425	-0.001	0.012	0.874
Extrinsic Motivator	425	-0.000	-0.075	0.323
Warmth	425	-0.117	-0.073	0.497
Classroom Characteristics				
Classroom size	425	30.774	30.597	0.739
Refugee Share	425	0.067	0.066	0.989

Note: The table presents the balance at baseline for the pooled sample. The p-values from the test of equality between control and treatment are shown in the last column. The p-value from joint test of student characteristics is 0.967. The p-value from joint test of teacher and classroom characteristics is 0.904. Test scores and survey items are standardized to have a mean zero and a standard deviation of 1.

Table A2—Multiple Hypothesis Testing

	Original P-Value	Sharpened Q-Value	Romano Wolf P-Value
Panel A: Student Outcomes			
Experimental Task			
Science Related Booklet	0.001	0.004	0.024
Non-Science Booklet	0.393	0.199	0.583
No Booklet	0.000	0.003	0.016
WTP(All)	0.007	0.015	0.048
WTP (Science)	0.000	0.002	0.014
WTP (Non-Science)	0.712	0.350	0.719
Retention	0.011	0.021	0.060
Science Retention	0.012	0.021	0.060
Non-Science Retention	0.027	0.028	0.092
Achievement & Aspirations			
Science	0.017	0.024	0.146
Math	0.645	0.326	0.944
Verbal	0.232	0.145	0.681
University Aspiration	0.086	0.064	0.411
Science Aspiration	0.001	0.003	0.018
Engineering Aspiration	0.944	0.447	0.962
Medical Aspiration	0.798	0.385	0.962
Non-STEM Aspiration	0.057	0.052	0.329
Students' Beliefs & Attitudes			
Grit	0.225	0.145	0.443
Impulsivity	0.548	0.281	0.623
Risk	0.079	0.062	0.261
Ambiguity	0.011	0.021	0.048
Critical Thinking	0.006	0.015	0.036
Curiosity Survey	0.000	0.001	0.002
Science Curiosity	0.000	0.001	0.002
Panel B: Teacher Outcomes			
Curiosity	0.002	0.011	0.026
Modern Teaching	0.046	0.135	0.349
Warmth	0.053	0.135	0.357
Extrinsic Motivator	0.302	0.734	0.914
Growth Mindset	0.000	0.005	0.010
Professional Attachment	0.813	1.000	0.990
Competence Beliefs	0.871	1.000	0.990
Gender Styping	0.399	0.823	0.942
Critical Thinking	0.758	1.000	0.990
Curricular Knowledge in Science	0.898	1.000	0.990
Booklet Knowledge	0.451	0.823	0.942

Note: The table presents estimation results for sharpened False Discovery Rate (FDR) q-values (Anderson, 2008) and adjusted p-values via Romano and Wolf (2005) multiple hypothesis correction. To accommodate Romano-Wolf correction to control for family wise error rate (FWER), we group our outcome variables into three, namely (i) experimental outcomes, (ii) achievement and aspiration related outcomes, (iii) beliefs and attitudes.

TABLE A3—HETEROGENEOUS TREATMENT EFFECTS - GENDER

Panel A: Choice of Booklet

	Science Related	Non-Science Related	No booklet
Treatment = Girls	0.077	-0.042	-0.035
	(0.017)	(0.017)	(0.010)
Treatment = Boys	0.000	0.023	-0.023
	(0.017)	(0.016)	(0.008)
P-Value : Girls=Boys	0.003	0.010	0.213
Control Mean - Girls	0.496	0.437	0.067
Control Mean - Boys	0.495	0.443	0.062
Observations	10898	10898	10898
Number of Schools	134	134	134

Panel B: Willingness to Pay

	WTP (All)	WTP (Science)	WTP (Non-Science)
Treatment = Girls	0.146	0.171	-0.054
	(0.047)	(0.034)	(0.033)
Treatment = Boys	0.073	0.026	0.035
	(0.040)	(0.034)	(0.037)
P-Value : Girls=Boys	0.057	0.001	0.068
Control Mean - Girls	-0.062	-0.010	-0.041
Control Mean - Boys	0.060	0.010	0.039
Observations	10892	10891	10891
Number of Schools	134	134	134

Note: Estimates are obtained via OLS. Panel A reports the estimated effects on the choice of a booklet. The dependent variables are binary indicators of choosing a science-related booklet (science, space, vehicles, human body, and animals) in column 1, choosing a nonscience-related booklet (history, sports, and cartoons) in column 2, and choosing no booklet option in column 3. Panel B reports estimated effects on the WTP for a booklet, WTP for a science-related booklet, and WTP for a non-science booklet. Covariates, selected via post-double-selection LASSO, include gender, fluid IQ, survey measure of curiosity, refugee status, math and verbal scores as individual baseline characteristics, class size, the share of refugees, teacher experience, and the number of children the teacher has. Grade and district fixed effects included. Grade and district fixed effects included. Standard errors are clustered at the school level and are reported in parentheses.

Table A4—Heterogeneous Treatment Effects - Gender

Panel A: Knowledge Retention

	Short Term			Long Term		
	Retention	Science Retention	Non-Science Retention	Retention	Science Retention	Non-Science Retention
Treatment = Girls	0.114	0.101	0.087	0.080	0.151	-0.047
	(0.056)	(0.050)	(0.050)	(0.077)	(0.064)	(0.084)
Treatment = Boys	0.114	0.103	0.085	0.193	0.161	0.160
	(0.054)	(0.049)	(0.049)	(0.096)	(0.082)	(0.093)
P-Value : Girls=Boys	0.997	0.966	0.965	0.362	0.920	0.096
Control Mean - Girls	-0.083	-0.079	-0.056	-0.031	-0.087	0.057
Control Mean - Boys	0.083	0.079	0.056	0.030	0.085	-0.055
Observations	9070	9070	9070	1335	1335	1335
Number of Schools	134	134	134	50	50	50

Panel B: Knowledge Retention (excluding Preferred Booklet)

	Short Term			Long Term		
	Retention	Science Retention	Non-Science Retention	Retention	Science Retention	Non-Science Retention
Treatment = Girls	0.107	0.086	0.097	0.086	0.123	0.001
	(0.059)	(0.052)	(0.052)	(0.079)	(0.063)	(0.096)
Treatment = Boys	0.131	0.123	0.096	0.253	0.185	0.240
	(0.056)	(0.051)	(0.046)	(0.095)	(0.084)	(0.089)
P-Value : Girls=Boys	0.604	0.384	0.993	0.190	0.563	0.059
Control Mean - Girls	-0.073	-0.078	-0.031	-0.037	-0.079	0.030
Control Mean - Boys	0.073	0.077	0.031	0.035	0.075	-0.029
Observations	8299	8299	8299	1218	1218	1218
Number of Schools	134	134	134	50	50	50

Note: Estimates are obtained via OLS using the sample restricted to the half-half regime. The dependent variables are standardized booklet test scores (knowledge retention). The first three columns give short-term results using the pooled sample, and the last three provide the long-term results of Study 1. Covariates for the short-term specification, selected via post-double-selection LASSO, include gender, fluid IQ, survey measure of curiosity, refugee status, math and verbal scores as individual baseline characteristics, class size, the share of refugees, teacher experience, and the number of children the teacher has. The long-term covariate set, selected via post-double-selection LASSO, is similar but excludes class size and refugee share. Grade and district fixed effects included. Standard errors are clustered at the school level and are reported in parentheses.

Table A5—Heterogeneous Treatment Effects - Gender

	Short Term			Long Term		
	Science	Math	Verbal	Science	Math	Verbal
Treatment = Girls	0.056	0.015	0.024	0.100	-0.031	0.005
	(0.040)	(0.031)	(0.029)	(0.050)	(0.049)	(0.055)
Treatment = Boys	0.091	0.011	0.040	0.045	-0.002	-0.016
	(0.033)	(0.030)	(0.032)	(0.064)	(0.055)	(0.063)
P-Value : Girls=Boys	0.384	0.856	0.583	0.469	0.639	0.761
Control Mean - Girls	-0.005	-0.021	0.114	-0.020	0.020	0.120
Control Mean - Boys	0.005	0.021	-0.113	0.020	-0.020	-0.118
Observations	9977	10433	10713	2424	2424	2424
Number of Schools	134	134	134	50	50	50

Note: Estimates are obtained via OLS. The dependent variables are standardized subject test scores. The first 3 columns give short-term results using the pooled sample, and the last 3 provide the long-term results of Study 1. Covariates for the short-term specification, selected via post-double-selection LASSO, include gender, fluid IQ, survey measure of curiosity, refugee status, math and verbal scores as individual baseline characteristics, class size, the share of refugees, teacher experience, and the number of children the teacher has. The long-term covariate set, selected via post-double-selection LASSO, is similar but excludes class size and refugee share. Grade and district fixed effects included. Standard errors are clustered at the school level and are reported in parentheses.

Table A6—Heterogeneous Treatment Effects - Gender

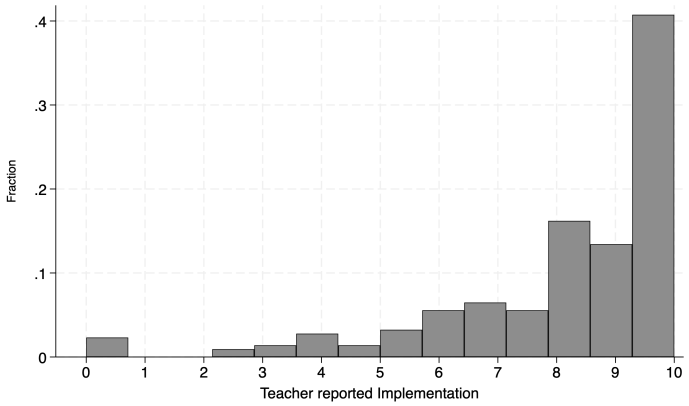
Panel A: Short Term

	University	Science	Engineering	Medical	Non-STEM
Treatment = Girls	0.009	0.030	0.002	-0.008	-0.024
	(0.006)	(0.008)	(0.008)	(0.013)	(0.015)
Treatment = Boys	0.008	0.015	-0.001	0.004	-0.018
	(0.007)	(0.011)	(0.011)	(0.009)	(0.016)
P-Value : Girls=Boys	0.913	0.300	0.791	0.411	0.782
Control Mean - Girls	0.963	0.079	0.063	0.228	0.630
Control Mean - Boys	0.937	0.151	0.174	0.095	0.580
Observations	10721	10212	10212	10212	10212
Number of Schools	134	134	134	134	134

Panel B: Long Term

	University	Science	Engineering	Medical	Non-STEM
Treatment = Girls	0.004	0.024	0.006	-0.040	0.011
	(0.010)	(0.025)	(0.019)	(0.026)	(0.038)
Treatment = Boys	0.013	0.000	0.022	0.018	-0.040
	(0.014)	(0.027)	(0.030)	(0.023)	(0.035)
P-Value : Girls=Boys	0.612	0.482	0.667	0.100	0.374
Control Mean - Girls	0.963	0.095	0.048	0.299	0.558
Control Mean - Boys	0.936	0.163	0.190	0.127	0.520
Observations	2318	2181	2181	2181	2181
Number of Schools	50	50	50	50	50

Note: Estimates are obtained via OLS. The dependent variables are binary choice variables of intention to go to university, intention to choose a science major, engineering major, medicine, and non-STEM major. Panel A presents short-term results from the pooled sample, and Panel B long-term results from Study 1. Covariates for the short-term specification, selected via post-double-selection LASSO, include gender, fluid IQ, survey measure of curiosity, refugee status, math and verbal scores as individual baseline characteristics, class size, the share of refugees, teacher experience, and the number of children the teacher has. The long-term covariate set, selected via post-double-selection LASSO, is similar but excludes class size and refugee share. Grade and district fixed effects included. Standard errors are clustered at the school level and are reported in parentheses.



Note: The figure depicts the program implementation intensity reported by treated teachers at endline. Teachers were given a 10cm line that has a moving cursor to report the level they believe represents their implementation intensity, zero representing no implementation, and 10 a 100% implementation.

FIGURE A1. IMPLEMENTATION INTENSITY

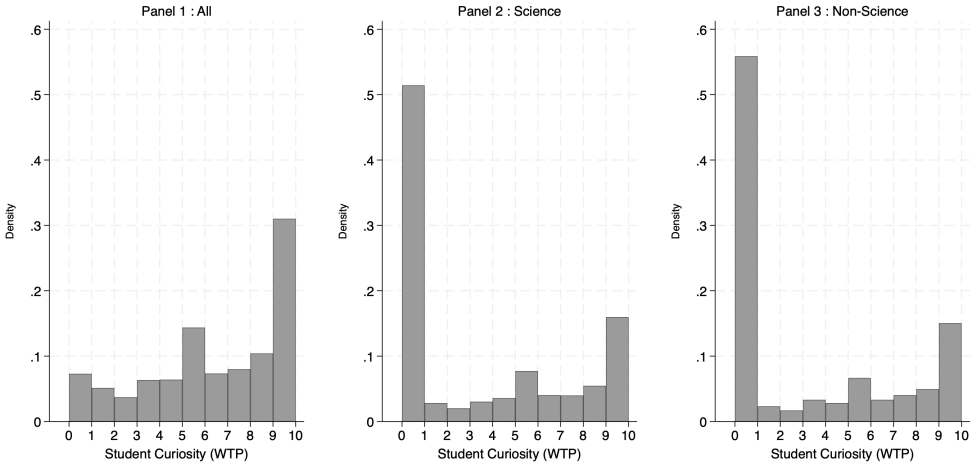


FIGURE A2. STUDENT CURIOSITY DISTRIBUTION (WTP)

Note: Figures depict the distribution of the number of tokens forgone for a booklet (Panel A), for a science-related booklet (Panel B), and for a non-science related booklet (Panel C).

*

REFERENCES

Abeberese, Ama Baafra, Todd J. Kumler, and Leigh L. Linden. 2014. "Improving Reading Skills by Encouraging Children to Read in School:: A Randomized Evaluation of the Sa Aklat Sisikat Reading Program in the Philippines." *Journal of Human Resources*, 49(3): 611–633.

Alan, Sule. 2021. "Stimulating Curiosity to Enhance Learning: Results from a Randomized Intervention-Replication Trial." AEA RCT Registry. November 30. https://doi.org/10.1257/rct.8629-1.0.

Alan, Sule, and Ipek Mumcu. 2023. "Stimulating Curiosity to Enhance Learning: Results from a Randomized Intervention." AEA RCT Registry. January 16. https://doi.org/10.1257/rct.3957-1.2.

Alan, Sule, and Seda Ertac. 2018. "Fostering Patience in the Classroom: Results from Randomized Educational Intervention." *Journal of Political Economy*, 126(5): 1865–1911.

Alan, Sule, Ceren Baysan, Mert Gumren, and Elif Kubilay. 2021. "Building Social Cohesion in Ethnically Mixed Schools: An Intervention on Perspective Taking*." *The Quarterly Journal of Economics*, 136(4): 2147–2194.

Alan, Sule, Teodora Boneva, and Seda 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.

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

ASER. 2018. "Annual Status of Education Report 'Beyond Basics' (Rural) 2017." ASER Centre.

Ashraf, Nava, Abhijit Banerjee, and Vesall Nourani. 2021. "Learning to Teach by Learning to Learn." 115.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukherji, Marc Shotland, and Michael Walton. 2016. "Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of "Teaching at the Right Level" in India." National Bureau of Economic Research Working Paper 22746.

Banerjee, Abhijit V., Shawn Cole, Esther Duflo, and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *The Quarterly Journal of Economics*, 122(3): 1235–1264.

Banerji, Rukmini, and Madhav Chavan. 2016. "Improving Literacy and Math Instruction at Scale in India's Primary Schools: The Case of Pratham's Read India Program." *Journal of Educational Change*, 17(4): 453–475.

Baron-Cohen, Simon, Therese Jolliffe, Catherine Mortimore, and Mary Robertson. 1997. "Another Advanced Test of Theory of Mind: Evidence from Very High Functioning Adults with Autism or Asperger Syndrome." *Journal of Child Psychology and Psychiatry*, 38(7): 813–822.

Berlyne, D. E. 1954. "A Theory of Human Curiosity." *British Journal of Psychology. General Section*, 45(3): 180–191.

Blanchard, Margaret R., Sherry A. Southerland, and Ellen M. Granger. 2009. "No Silver Bullet for Inquiry: Making Sense of Teacher Change Following an Inquiry-Based Research Experience for Teachers." *Science Education*, 93(2): 322–360.

Brown, Christina L., Supreet Kaur, Geeta Kingdon, and Heather Schofield. 2022. "Cognitive Endurance as Human Capital."

Buser, Thomas, Noemi Peter, and Stefan C. Wolter. 2017. "Gender, Competitiveness, and Study Choices in High School: Evidence from Switzerland." *American Economic Review*, 107(5): 125–130.

Carlana, Michela, and Margherita Fort. 2022. "Hacking Gender Stereotypes: Girls' Participation in Coding Clubs." *AEA Papers and Proceedings*, 112: 583–587.

Chamorro-Premuzic, Tomas, and Adrian Furnham. 2006. "Intellectual Competence and the Intelligent Personality: A Third Way in Differential Psychology:." *Review of General Psychology*.

Collins, Robert P, Jordan A Litman, and Charles D Spielberger. 2004. "The Measurement of Perceptual Curiosity." *Personality and Individual Differences*, 36(5): 1127–1141.

de Ree, Joppe, Karthik Muralidharan, Menno Pradhan, and Halsey Rogers. 2018. "Double for Nothing? Experimental Evidence on an Unconditional Teacher Salary Increase in Indonesia." *The Quarterly Journal of Economics*, 133(2): 993–1039.

Dubey, Rachit, Hermish Mehta, and Tania Lombrozo. 2021. "Curiosity Is Contagious: A Social Influence Intervention to Induce Curiosity." *Cognitive Science*, 45(2): e12937.

Duckworth, Angela Lee, and Patrick D. Quinn. 2009. "Development and Validation of the Short Grit Scale (Grit–S)." *Journal of Personality Assessment*, 91(2): 166–174.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." *American Economic Review*, 101(5): 1739–1774.

Dweck, Carol S. 2008. *Mindset: The New Psychology of Success.* Random House Digital, Inc.

Fischer, Stefanie. 2017. "The Downside of Good Peers: How Classroom Composition Differentially Affects Men's and Women's STEM Persistence." *Labour Economics*, 46: 211–226.

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.* Vol. 5, , ed. Eric A. Hanushek, Stephen Machin and Ludger Woessmann, 653–743. Elsevier.

Glewwe, Paul, Michael Kremer, Sylvie Moulin, and Eric Zitzewitz. 2004. "Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya." *Journal of Development Economics*, 74(1): 251–268.

Glewwe, Paul, Sylvie Lambert, and Qihui Chen. 2020. "Chapter 15 - Education Production Functions: Updated Evidence from Developing Countries." In *The Economics of Education (Second Edition).* , ed. Steve Bradley and Colin Green, 183–215. Academic Press.

Gneezy, Uri, and Jan Potters. 1997. "An Experiment on Risk Taking and Evaluation Periods*." *The Quarterly Journal of Economics*, 112(2): 631–645.

Goldhaber, Dan, Thomas J. Kane, Andrew McEachin, Emily Morton, Tyler Patterson, and Douglas O. Staiger. 2022. "The Consequences of Remote and Hybrid Instruction During the Pandemic." National Bureau of Economic Research, Inc 30010.

Gottfried, Adele Eskeles, Kathleen Suzanne Johnson Preston, Allen W. Gottfried, Pamella H. Oliver, Danielle E. Delany, and Sirena M. Ibrahim. 2016. "Pathways from Parental Stimulation of Children's Curiosity to High School Science Course Accomplishments and Science Career Interest and Skill." *International Journal of Science Education*, 38(12): 1972–1995.

Granger, E. M., T. H. Bevis, Y. Saka, S. A. Southerland, V. Sampson, and R. L. Tate. 2012. "The Efficacy of Student-Centered Instruction in Supporting Science Learning." *Science*, 338(6103): 105–108.

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

Gruber, Matthias J., and Charan Ranganath. 2019. "How Curiosity Enhances Hippocampus-Dependent Memory: The Prediction, Appraisal, Curiosity, and Exploration (PACE) Framework." *Trends in Cognitive Sciences*, 23(12): 1014–1025.

Gruber, Matthias J., Bernard D. Gelman, and Charan Ranganath. 2014. "States of Curiosity Modulate Hippocampus-Dependent Learning via the Dopaminergic Circuit." *Neuron*, 84(2): 486–496.

Gust, Sarah, Eric A. Hanushek, and Ludger Woessmann. 2022. "Global Universal Basic Skills: Current Deficits and Implications for World Development."

Hartung, Freda-Marie, and Britta Renner. 2013. "Social Curiosity and Gossip: Related but Different Drives of Social Functioning." *PLOS ONE*, 8(7): e69996.

Hjort, Jonas, Diana Moreira, Gautam Rao, and Juan Francisco Santini. 2021. "How Research Affects Policy: Experimental Evidence from 2,150 Brazilian Municipalities." *American Economic Review*, 111(5): 1442–1480.

James, William. 1983. *Talks to Teachers on Psychology and to Students on Some of Life's Ideals.* Harvard University Press.

Jirout, Jamie, and David Klahr. 2012. "Children's Scientific Curiosity: In Search of an Operational Definition of an Elusive Concept." *Developmental Review*, 32(2): 125–160.

Kahn, Shulamit, and Donna Ginther. 2017. "Women and STEM."

Kashdan, Todd B., and Paul J. Silvia. 2009. "Curiosity and Interest: The Benefits of Thriving on Novelty and Challenge." In *Oxford Handbook of Positive Psychology, 2nd Ed. Oxford Library of Psychology*, 367–374. New York, NY, US:Oxford University Press.

Kashdan, Todd B., David J. Disabato, Fallon R. Goodman, and Patrick E. McKnight. 2020. "The Five-Dimensional Curiosity Scale Revised (5DCR): Briefer Subscales While Separating Overt and Covert Social Curiosity." *Personality and Individual Differences*, 157: 109836.

Kremer, Michael, Conner Brannen, and Rachel Glennerster. 2013. "The Challenge of Education and Learning in the Developing World." *Science*, 340(6130): 297–300.

Kremer, Michael, Paul Glewwe, and Sylvie Moulin. 2009. "Many Children Left Behind? Textbooks and Test Scores in Kenya." *American Economic Journal: Applied Economics*, 1(1 (January 2009)): 112–135.

Litman, Jordan A., and Charles D. Spielberger. 2003. "Measuring Epistemic Curiosity and Its Diversive and Specific Components." *Journal of Personality Assessment*, 80(1): 75–86.

Litman, Jordan A., and Mark V. Pezzo. 2007. "Dimensionality of Interpersonal Curiosity." *Personality and Individual Differences*, 43(6): 1448–1459.

Litman, Jordan A., Robert P. Collins, and Charles D. Spielberger. 2005. "The Nature and Measurement of Sensory Curiosity." *Personality and Individual Differences*, 39(6): 1123–1133.

Loewenstein, George. 1994. "The Psychology of Curiosity: A Review and Reinterpretation." *Psychological Bulletin*, 116(1): 75–98.

Majerowicz, Stephanie, and Ricardo Montero. 2022. "Can Teaching Be Taught? Experimental Evidence from a Teacher Coaching Program in Peru (Job Market Paper)."

OECD. 2013. "TALIS User Guide for the International Database."

Popova, Anna, David K Evans, Mary E Breeding, and Violeta Arancibia. 2022. "Teacher Professional Development around the World: The Gap between Evidence and Practice." *The World Bank Research Observer*, 37(1): 107–136.

Raven, John C, and John Hugh Court. 1998. *Raven's Progressive Matrices and Vocabulary Scales.* Vol. 759, Oxford pyschologists Press Oxford.

Romano, Joseph P, and Michael Wolf. 2005. "Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing." *Journal of the American Statistical Association*, 100(469): 94–108.

Shah, Prachi E., Heidi M. Weeks, Blair Richards, and Niko Kaciroti. 2018. "Early Childhood Curiosity and Kindergarten Reading and Math Academic Achievement." *Pediatric Research*, 84(3): 380–386.

Singh, Abhijeet, Mauricio Romero, and Karthik Muralidharan. 2022. "Covid-19 Learning Loss and Recovery: Panel Data Evidence from India."

Sleddens, Ester F. C., Stef P. J. Kremers, Nanne K. De Vries, and Carel Thijs. 2013. "Measuring Child Temperament: Validation of a 3-item Temperament Measure and 13-item Impulsivity Scale." *European Journal of Developmental Psychology*, 10(3): 392–401.

Sosu, Edward M. 2013. "The Development and Psychometric Validation of a Critical Thinking Disposition Scale." *Thinking Skills and Creativity*, 9: 107–119.

Terrenghi, Ilaria, Barbara Diana, Valentino Zurloni, Pier Cesare Rivoltella, Massimiliano Elia, Marta Castañer, Oleguer Camerino, and M. Teresa Anguera. 2019. "Episode of Situated Learning to Enhance Student Engagement and Promote Deep Learning: Preliminary Results in a High School Classroom." *Frontiers in Psychology*, 10.

Vogl, Elisabeth, Reinhard Pekrun, Kou Murayama, and Kristina Loderer. 2019*a*. "Surprised–Curious–Confused: Epistemic Emotions and Knowledge Exploration." *Emotion*, No Pagination Specified–No Pagination Specified.

Vogl, Elisabeth, Reinhard Pekrun, Kou Murayama, Kristina Loderer, and Sandra Schubert. 2019*b*. "Surprise, Curiosity, and Confusion Promote Knowledge Exploration: Evidence for Robust Effects of Epistemic Emotions." *Frontiers in Psychology*, 10.

von Stumm, Sophie, Benedikt Hell, and Tomas Chamorro-Premuzic. 2011. "The Hungry Mind: Intellectual Curiosity Is the Third Pillar of Academic Performance." *Perspectives on Psychological Science*, 6(6): 574–588.

World Bank, FCDO, and BE2. 2020. "Cost-Effective Approaches to Improve Global Learning: What Does Recent Evidence Tell Us Are "Smart Buys" for Improving Learning in Low- and Middle-Income Countries? Recommendations of the Global Education Evidence Advisory Panel." World Bank, FCDO, BE2.