Fighting the Learning Crisis in Developing Countries: A Randomized Experiment of Self-Learning at the Right Level

YASUYUKI SAWADA
University of Tokyo

MINHAJ MAHMUD
Asian Development Bank

MAI SEKI
Ritsumeikan University

HIKARU KAWARAZAKI
University College London and the Institute for Fiscal Studies

I. Introduction

Learning crisis refers to the global phenomenon in which more than 60% of children who complete their primary education in low- and middle-income countries fail to achieve a minimum proficiency in math and reading (UNESCO

This is a substantially revised version of the paper earlier circulated with the title "Individualized Self-Learning Program to Improve Primary Education: Evidence from a Randomized Field Experiment in Bangladesh." The opinions expressed in this paper are our own and do not reflect the views of affiliated organizations. We thank An Le, Saori Nishimura, and Kazuma Takakura for superb research assistance. We are grateful to the editor, Marcel Fafchamps, an associate editor, and two anonymous referees of the journal for their constructive comments. We also thank Esther Duflo, Pascaline Dupas, Deon Filmer, Dean Karlan, Halsey Rogers, and Paul Romer. Furthermore, we thank the session participants at the American Economic Association Meeting 2018, World Bank Education Global Practice Seminar, the 2017 European, North American Summer, and Asian Meetings of the Econometric Society, the Australasian Development Economics Workshop 2017, the Midwest International Economic Development Conference 2017, the National Graduate Institute for Policy Studies (GRIPS) and University of Tokyo Joint Workshop 2017, Hitotsubashi University, Kansai Labor Economics Workshop, and the Hayami Conference 2016 for their useful comments. We are grateful to the authorities of the BRAC; Kumon Institute of Education Co., Ltd.; and the Japan International Cooperation Agency for their cooperation in implementing the study. The study protocol went through a full committee review and was approved by the institutional review board of the University of Tokyo (ref. no. 15–90). The study has been registered at the American Economic Association's Randomized Controlled Trials Registry (no. AEARCTR-0002925). This work was supported by a grant

Electronically published June 11, 2024

Economic Development and Cultural Change, volume 72, number 4, July 2024.

https://doi.org/10.1086/725909

2013; World Bank 2018). Furthermore, improving the quality of education is a sine qua non for achieving the United Nations' Sustainable Development Goals (United Nations 2018). Owing to their high effectiveness in improving learning outcomes, teaching at the right level (TaRL) programs are gaining attention (Banerjee et al. 2007, 2016; Duflo, Dupas, and Kremer 2011; Muralidharan, Singh, and Ganimian 2019).[1] For example, Muralidharan, Singh, and Ganimian (2019) find that individualized technology-aided instruction programs in India can improve test scores. However, the lack of appropriate infrastructure in developing countries potentially constrains the use of such effective programs.

In this study, we evaluate the effectiveness of an individualized self-learning program in Bangladesh, the Kumon method of learning (hereafter, Kumon), which is based on the paper-and-pencil method and does not necessarily rely on the use of information and communication technology (ICT) in supplementing the learning quality of primary schools. Kumon is a globally popular, nonformal education program designed to ensure that each student always studies at the level that is "just right" for him or her.[2] In Kumon, each student begins at an individually suitable starting point identified through a diagnostic test and learns new concepts in small steps wherein learning is enforced through easily understandable hints and examples.

Bangladesh has successfully increased school enrollment and narrowed gender gaps. In addition to conventional public formal education, nonformal education has been critical to this process. In this respect, nongovernmental organizations (NGOs), such as BRAC, have played an important role in collaboration with the government. In particular, BRAC primary schools (BPSs) have provided disadvantaged students with a 4-year accelerated program that covers the 5-year public primary school curriculum.[3] Given the success of BPSs in ensuring

from Grants-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (no. KAKENHI 26220502). We declare that we have no financial interest related to this study. We designed and conducted an RCT evaluation independent of Kumon Institute of Education Co., Ltd. Data are provided through Dataverse at https://doi.org/10.7910/DVN/OADRDM. Contact the corresponding author, Yasuyuki Sawada, at sawada@e.u-tokyo.ac.jp.

[1] Regarding improving learning outcomes, demand-side approaches seem less promising than supply-side interventions (e.g., increasing the numbers of teachers and schools). See Asim et al. (2017) for a meta-analysis of effect evaluation studies that focus on improving learning outcomes in South Asian countries. Other reviews that focus on the effects of interventions on learning outcomes include Kremer, Brannen, and Glennerster (2013); Glewwe (2014); Evans and Popova (2015); McEwan (2015); and Ganimian and Murnane (2016).

[2] As of March 2024, there were 3.55 million Kumon subject enrollments, and the program had been adopted in 63 countries and regions. See https://www.kumongroup.com/eng/about/ for details.

[3] BPSs are regarded as one of the largest and most successful nonformal education programs that are targeted at disadvantaged populations in Bangladesh. They have introduced a seasonally adjusted school

enrollment and reducing the number of primary school dropouts, the Bangladesh government has scaled up a modified version of BPS under the Reaching Out of School project; the goal is to provide a low-cost platform to target children from difficult-to-reach communities who are out of school (Asadullah 2016). Despite these efforts, a lack of quality education and resulting inadequate student learning remain a serious concern in the country, as in other developing countries.[4]

In this context, we adopt and evaluate the effect of Kumon in improving both cognitive and noncognitive abilities of BPS students in Bangladesh, given Kumon's unique setting in providing nonformal education and internal efficiency, unlike formal schools (Ahmad and Haque 2011). While Kumon is a globally popular supplementary education method in improving both cognitive and noncognitive abilities, our study is, to our knowledge, the first to experimentally investigate its effect on these abilities. BPSs have 30 students per class with diverse backgrounds and a large variance in terms of ability in the subjects taught, particularly math (Nath 2012). This creates a potential mismatch between the teaching level and students' individual abilities. However, BPSs cannot effectively offer TaRL, as they follow the same instructional approach as that used in public schools. Kumon, as a supplementary approach, could at least partially respond to this mismatch and improve learning outcomes by providing self-learning math materials for each student.[5] Indeed, Kumon's goal has been to improve both cognitive ability and certain noncognitive abilities (e.g., perceived competence, self-confidence, and self-esteem). According to the website of the Kumon Institute of Education, "As students take on the challenge of studying new material, they improve their concentration, learn to take on new challenges, develop perseverance, and gain a positive

calendar, which has been key to their success (Watkins 2000; Chowdhury, Jenkins, and Nandita 2014). Section II provides more details about BPSs.

[4] For example, Asadullah and Chaudhury (2013) find an imperfect correlation between years of schooling and cognitive outcomes: among children who completed primary schooling, only 49% could provide 75% or more correct answers in a simple arithmetic test, and the likelihood of providing more than 75% correct answers was only 9 percentage points higher than the likelihood for those with no schooling at all.

[5] While many existing studies have established the link between measured cognitive ability (e.g., IQ) and educational outcomes (e.g., schooling attainment and wages), recent studies have begun to shed new light on the role of noncognitive abilities (e.g., personality traits, motivations, and preferences; Heckman 2006, 2007). In fact, recent studies show that the predictive power of noncognitive abilities is comparable with or exceeds that of cognitive skills in terms of explaining education, success in the labor market, or other outcomes (Heckman 2006; Heckman, Humphries, and Kautz 2014). Because Kumon has been regarded as a successful nonformal education program in developed countries, its effects on learning outcomes in a disadvantaged setting in a developing country context are worth evaluating.

sense of self."[6] Therefore, improvements in cognitive abilities are expected to result through a "building block" of development of noncognitive outcomes. During Kumon sessions, all students have to concentrate on their chosen subject for 30 minutes every day. This technique would help develop noncognitive ability even among students who are initially lagging in their cognitive ability. In this manner, the Kumon intervention first improves the noncognitive ability of students who are initially lagging in both cognitive and noncognitive abilities. It should be noted that compared with commercially operated Kumon centers elsewhere, the deployed resources are generally limited in the Kumon program run in BPSs. Although we followed the standard Kumon worksheets, protocol, and routine procedures, we did not require students to complete any homework. In addition, unlike the standard Kumon centers, which offer sessions outside schools, our treatment-school students attended the Kumon session in the classroom prior to their regular classes.

We measure the cognitive ability improvements by comparing the math test scores obtained by students at the baseline and end line, known as the diagnostic test (DT) score. The findings indicate that Kumon substantially improves students' cognitive abilities as measured by the DT score and the effect size of 0.465 SD.[7] Given that our intervention is designed to increase students' math problem-solving skills in a time-efficient manner, we show the effect using test scores per minute wherein the effect comes through both test score gains and reductions in the problem-solving speed. Therefore, the magnitude of the effect through test score per minute (2.085 SD) is much higher than the effect size of education interventions elsewhere. Notably, this is largely due to a substantial reduction in test completion time, as reflected in the effect size of the DT time (−2.209 SD). To the best of our knowledge, this is the first study in the economics literature to employ a time-adjusted test outcome as a critical measure of cognitive ability, consistent with the educational and psychological literature (AERA, APA, and NCME 2014; Engelhardt and Goldhammer 2019). Additionally, we measure cognitive ability using a second math test score, which is known as the Proficiency Tests of Self-Learning Skills (PTSII-C) score. The PTSII-C not only captures accuracy but also tests how many problems students could attempt to solve within the specified time. In other words, their score already reflects both accuracy and speed. Indeed, the effect size in the case of the

[6] See https://www.kumongroup.com/eng/about-kumon/future/ for details.

[7] These effects are largely comparable with some existing interventions. For example, Lakshminarayana et al. (2013) find a 0.75 SD effect from the supplementary remedial teaching provided by Indian NGOs on students' test scores in public primary schools. Furthermore, Duflo, Dupas, and Kremer (2011) find a 0.9 SD effect from the peer effects of tracking for the top quantile of students in Kenyan primary schools.

PTSII-C score is comparably high (i.e., 0.999 SD). Regarding noncognitive abilities measured through certain personality traits, we find catch-up effects among students who have initially lower abilities compared with those of the median.

We also show a longer-term effect of the intervention using these students' academic achievements in the national-level Primary School Certificate (PSC) examination held 8 months (grade 4 students) and 20 months (grade 3 students) after the intervention. In particular, we measure students' development in math ability by comparing their PSC math score and the baseline PTSII-C score. We apply quasi-experimental and bounds analyses to address potential sample-selection issues. Overall, we find a modest but positive long-term effect of the intervention on cognitive ability; the average treatment effects range between 0.233 and 0.235 SD, which is within Lee's treatment effect bounds (Lee 2009). Additionally, we show that the cost exceeds the benefit under some reasonable assumptions, and Kumon could be a cost-effective complementary intervention to existing lecture-style, primary education curricula.

The remainder of this paper is organized as follows. In section II, we outline our experimental design, including the setting and the intervention, and then explain the data and baseline test results. Section III presents the econometric evaluation framework, followed by the empirical results. Section IV compares the benefits and costs of this intervention. Finally, section V discusses the findings and limitations.

II. Experimental Design, Data, and Balancing Test

A. Setting: BRAC Primary School

Primarily, BPS targets children from disadvantaged social backgrounds who could not access formal schooling at the right age or have dropped out of the formal education system. The economic eligibility criteria stipulate that "children of poor households having less than 50 decimals of land and at least one member of the household that has worked for wages for at least 100 days" and those who are living within a 2 km radius of the school are admitted to BPS (Afroze 2012, 1). BPS covers the same standard curriculum as public schools. Although BPS and government primary schools teach the same competency-based curriculum, they have some basic differences. Unlike the 5-year standard primary school system, BPS offers an accelerated 4-year program to help these children readapt to formal education (Asadullah 2016). In particular, BPS teachers address students who are falling behind in the following manner: the entry age for students in BPS is higher than that in standard primary schools (the official age is 6 years for entry into primary education); the schools operate under a rather flexible time schedule for 3 hours a day, 6 days a week, with fewer

holidays than government schools, resulting in higher contact hours per primary cycle than government primary schools. The average class size in BPS (i.e., 25–30 students) is smaller than that of government primary schools. BPSs are essentially one-classroom, one-teacher schools, and the teacher teaches all subjects to the same cohort. However, the pedagogical approach is influenced by traditional methods, such as group lectures followed by assignments. Students are required to pass the grade 5 terminal examination set by the government (i.e., the PSC exam). This also suggests that BPS provides learners with the same skills that are taught in government schools; that is, teaching for the test potentially affects students' learning.

Thus, the Kumon intervention aims to promote self-learning by encouraging each student to study at the right level and to learn to set goals and take up challenges at the next level. Given the unique setting of this nonformal education (e.g., the low-cost platform and smaller class size), BPS has the potential to scale up this intervention to supplement learning quality in primary education in Bangladesh by developing students' cognitive and noncognitive abilities.

B. Intervention: The Kumon Method of Learning

As a supplementary module in math, the Kumon method of learning has been introduced in selected BPSs among grade 3 and grade 4 students.

1. The Kumon Method

In general, Kumon aims to enable students to develop advanced academic and self-learning abilities by ensuring that they always study at a level that is appropriate for them. Students are assigned to an initial level on the basis of their individual performance in a DT rather than their school grade or age. The Kumon method is uniquely designed so that the initial level is slightly lower than a student's concurrent maximum capacity. This is for the following reasons: (i) to ensure that students fully understand the basic concepts and develop a firm foundation for the development of their cognitive abilities and (ii) to motivate them to continue studying, which also aids the development of their noncognitive abilities (e.g., self-esteem and sense of competence). Kumon worksheets, ranging from simple counting to advanced math, are designed with the level of difficulty increasing gradually. The worksheets contain example questions with hints and graphical explanations that help students independently acquire step-by-step problem-solving skills by themselves, not necessarily requiring high-level literacy.[8] Kumon instructors do not conduct

[8] See example worksheets of Kumon in figs. A1 and A2 of app. A (apps. A–J are available online).

lectures; they simply observe students' progress. They adjust the level of the worksheets if the students are stuck on the same worksheet or are unable to find the right answer after many attempts. Consequently, they can absorb materials beyond their school grade level through self-learning and advance to high-school-level materials at an early age. Importantly, slower learners can spend more time on basics without being rushed on to advanced-level materials beyond their level of understanding.

Another feature of Kumon is that it tracks each student's progress and achievements by use of personalized grade record books (hereafter, record books). Kumon instructors do not teach in class. Hence, they do not require extensive prior experience in conducting daily quizzes to monitor each student's understanding and progress. This is because Kumon worksheets are presented in small steps that enable students to learn independently by themselves. Furthermore, a set standard time is allocated to solve each worksheet, allowing BPS teachers to mechanically determine the level that the students are permitted to advance to or whether they should repeat a level. Detailed progress reports on the worksheets allow instructors to obtain more objective information about their students' abilities and understanding of the math involved.

2. Intervention in BPSs

Our intervention was a pilot program in BPSs to examine the effectiveness of Kumon in a disadvantageous setting subject to resource constraints and run during regular school hours. Unlike the regular Kumon sessions elsewhere, BPS provided a 30-minute Kumon session daily without any homework assignments. The learning materials were supplied by the Kumon Institute of Education Co., Ltd., of Japan after translating them into the local language (i.e., Bengali). The Kumon Institute also supplied training sessions for BPS teachers, who would supervise the Kumon sessions in BPSs. During Kumon sessions, the BPS teachers conducted no lectures. Instead, they observed their students' progress on individualized worksheets without intervention in principle. When a student became stuck solving problems after many attempts, the BPS teachers adjusted the level of worksheets downward to facilitate individual learning according to the pre-fixed procedure they learned during the training sessions.[9] The BPS teachers were not responsible for grading or recording the

[9] There were short conversations between a BPS teacher and each student, but there was no direct teaching during the Kumon session. Additionally, the teachers needed to determine students' worksheet levels fairly mechanically on the basis of the scores and time in principle, as trained. Therefore, these interactions should have played no—if any—important role in students' learning.

marks. The designated marking assistants gave grades and recorded the marks in the prescribed record books. The grading assistants had a few hours of training on how to grade before the intervention and on-the-job training. Until the session ended, students either moved on to a new worksheet once they had achieved a full score on the previous worksheet or continued to attempt and correct their answers until they achieved a full score within the designated time frame. On rare occasions when students encountered great difficulty with higher-order problem-solving tasks beyond their grade level, the BPS teachers might have assisted only to clarify the examples in the worksheet.

C. Experimental Design

To identify the causal effects of Kumon on young students' learning and particularly their cognitive abilities, we designed and conducted a randomized controlled trial (RCT) evaluation. Consistent with the effect size of education intervention elsewhere, we hypothesized a minimum detectable effect of 0.40 SD on students' cognitive ability. In our context, we referred to the results from studies of high-impact education interventions that involved TaRL, such as those of Lakshminarayana et al. (2013; 0.75 SD) and Duflo, Dupas, and Kremer (2011; 0.9 SD), and hypothesized the effect size to be 0.4 SD. Considering that randomization is conducted at the cluster (school/classroom) level, we assumed an intracluster correlation of 0.10 and a statistical significance of less than .05 for a two-tailed test. Thus, a sample of approximately 26 clusters with a statistical power of 0.80 was obtained. To ensure that we did not lose statistical power because of attrition or other factors, we selected a cluster size of 34 to increase the total student sample and an average of 30 students per cluster. This gave us a final sample of approximately 1,000 students. Then, we randomly selected 34 schools from a list of 179 eligible BPSs (located in Dhaka and surrounding areas) for our study, dividing them equally into 17 treatment and 17 control schools. The resulting sample breakdown by class/grade was as follows: 19 (out of 48 schools) for grade 3 and 15 (out of 131 schools) for grade 4.[10] The schools did not overlap in terms of grade. In other words, in a particular school, we offered the intervention only to grade 3 or to grade 4.

The intervention consisted of a 30-minute session on the Kumon method prior to the students' regular lessons. Thus, during the study period, the students in the treatment schools arrived at school earlier than their usual school

[10] On the basis of a complete list of 179 schools in Dhaka and nearby districts provided by BRAC, we randomly sampled schools by setting grade-specific strata. Accordingly, we randomly chose 18 and 16 for grade 3 and grade 4, respectively. One school listed for grade 4 turned out to be for grade 3, resulting in odd numbers of schools for each grade.

hours. Unlike the regular Kumon sessions elsewhere, we did not require students to complete related homework to restrict the daily 30-minute regular Kumon learning sessions. In addition, unlike a standard Kumon center that offers sessions outside school, our treatment-school students remained in the classroom where their regular BPS classes were held. BPSs run for 6 days a week, except on public holidays, teacher refreshment days, and teacher training days. Our intervention lasted for 8 months, from August 2015 to April 2016.

For the treatment schools, the Kumon Institute provided an intervention package comprising a math materials set and an instructor manual with sheets for the BRAC teachers.[11] The full materials set comprises (i) math worksheets with questions at various difficulty levels and achievement tests at the end of each level and (ii) a record book to track the students' daily progress. This included the level of worksheet that a student worked on, the number of repetitions required before achieving a full score on the worksheet, and the number of worksheets that students finally completed (fig. A3; figs. A1–I2 are in the online appendixes)).[12] We believe that our intervention was not necessarily ideal but sufficiently well designed to follow regular channels in a classroom setting without ICT infrastructure, and the results obtained were generalizable in the case of other intended beneficiaries in a similar setting. Although we followed the standard Kumon worksheets, protocol, and routine procedures, the deployed resources were generally limited compared with commercially operated Kumon centers elsewhere.

D. Data Description

We construct cognitive ability measures at both the baseline and end line on the basis of two different math test scores for both the treatment-school and control-school students. These math tests are the DT and the PTSII-C. The DT measures cognitive (math) abilities, whereby we retain records of both the score and time taken to complete the test.[13] The DT used for this study is time

[11] BRAC field staff were assigned to assist and follow up on BPS teachers. Prior to launching the program, a 3-day preparatory training for BPS teachers and field staff was held to familiarize them with the concepts and procedures of the learning method, followed by three additional 1-day training sessions during the intervention. Two marking assistants (graders) were provided for each class to support the grading and recording of the worksheets during Kumon sessions. The BPS teachers monitored the students and determined the level of worksheets that they were required to work on.

[12] All the materials, including numbers, were provided in the Bengali language, which was the medium of instruction for BPS teachers and students.

[13] In the standard assessment methods, time is one of the fundamental dimensions when constructing a test (AERA, APA, and NCME 2014), and the time information captures cognitive ability of a test taker. For example, nowadays time spent by test takers in each question is readily available in the case of computer-based test results. According to Engelhardt and Goldhammer (2019), time reflects

specific and requires students to answer 70 questions within a maximum of 10 minutes.[14] Hence, for the DT, we show the test scores per minute (DT score per minute) to determine the students' cognitive abilities. Meanwhile, the PTSII has two sections: The first section contains a total of 228 math questions within five categories that measure different dimensions of math problem-solving skills; here, the aggregate score defines students' cognitive ability (i.e., PTSII-C).[15] While the DT is a standard test wherein students are expected to complete all the questions in a given time frame, the PTSII-C test does not require the same. Instead, PTSII-C is designed in a manner that students answer as many questions as possible within a given time frame. However, they are not required to complete all the questions. PTSII-C not only captures the accuracy but also tests how many problems students attempt to solve within a specified time. The second section comprises 27 questions that measure the aspects of noncognitive abilities (see table C1; tables C1–J4 are in the online appendixes). Among the 27 questions, 8 are consistent with the Rosenberg Self-Esteem Scale (RSES index; Rosenberg 1965), and 10 are consistent with the Children's Perceived Competence Scale (CPCS index; Harter 1979; Sakurai and Matsui 1992). As noncognitive ability measures, we create the RSES and CPCS indexes on the basis of these questions.[16]

To assess the possible long-term effect of the intervention, we also collected students' results from the PSC exam, a nationally administered primary education completion test administered by the Ministry of Primary and Mass

duration of the cognitive process and thus can be considered in relation to the outcomes of the cognitive processing, implying that the time-adjusted test score is an indicator of cognitive ability.

[14] Although some time mismanagement occurred during the baseline DT (fig. B1), these cases are very few, and it is not likely that the time-reduction effects are entirely driven by these cases. Furthermore, timekeeping was strictly maintained in the end line across both treatment and control. As indicated in fig. B2, there are no observations going beyond the 10-minute limit.

[15] The PTSII-C includes 348 questions, which include 120 extremely simple tasks (part 1) and 228 simple math questions (parts 2–6). The former task questions asked students to connect the dots to form an alphabet to bring their focus and energy into problem solving. Part 1 was not used in the BPS; therefore, we do not use this in our analysis. Subsequently, the students were given 228 simple math questions: 80 quite simple addition and subtraction problems (part 2), 60 slightly difficult addition and subtraction problems (part 3), 28 problems for identifying a particular number from a sequence (part 4), 40 problems confirming answers to given addition and subtraction problems (part 5), and 20 questions filling the (blank) number in a sequence or in an addition or subtraction equation (part 6). Parts 2 and 3 are standard calculation problems found in any calculation problem set. These are similar to the DT and everyday worksheets. Parts 4 and 5 are unique to the PTSII, and students do not see these types in either the DT or everyday worksheets. Part 6 comprises a unique style of problems not commonly seen in standard calculation problems. However, these types of questions overlap with some parts of everyday worksheets.

[16] We adopt a short version of the RSES index, which is widely used in existing studies, including the study by Heckman, Stixrud, and Urzua (2006).

Education.[17] We particularly focus on PSC math results, given that our intervention was related to math problem-solving skills. Grade 4 (and grade 3) students had a chance to take the PSC exam for about 8 months (and 20 months) after the end of the intervention in December 2016 (and December 2017).[18]

We also conducted a teacher survey that captured teachers' assessments of students' performance. We collected each teacher's subjective evaluation of the performance of individual students at both the baseline and end line. Specifically, we asked each teacher about each student's performance through a 5-point Likert scale (very good; good; average; bad; and very bad). We then took the absolute distance between teachers' evaluations and observed test scores (i.e., DT or PTSII-C scores).

E. Balancing the Test Results

Baseline balance tests are performed by comparing the main variables of interest between the students of the treatment and control groups in addition to demographic variables. These include DT score, DT time, DT score per minute, PTSII-C score, variables measuring noncognitive abilities (i.e., RSES index and CPCS index), and students' characteristics (e.g., gender, age, and age squared). The mean and standard deviation of all raw scores of those who had end-line records of each variable are reported in table 1.[19] No significant differences were observed in the average baseline scores between the students of the treatment and control groups (baseline balance), suggesting the success of randomization.[20] It should be noted that the number of observations is smaller than the intended

[17] Those who wish to pursue further education must pass this exam. On the basis of the exam results, letter grades from A+ to A, A−, B, C, D, and F are assigned: if the score is in the range 80–100, the letter grade is an A+; if 70–79, it is an A; if 60–69, it is an A−; if 50–59, it is a B; if 40–49, it is a C; if 33–39, it is a D; and if below 33, it is an F. The subjects include, among others, Math and English. Unfortunately, we have no data on the exact score for an individual subject, but we do have data on the letter grades. See http://www.educationboard.gov.bd/computer/grading_system.php for details.

[18] Generally, this exam is administered at the end of grade 5 as a primary school terminal examination. As BPS adopts an accelerated curriculum that covers primary school requirements in grade 4, the students were allowed to take the PSC exam at the end of grade 4.

[19] The sample is for the analysis of covariance (ANCOVA) specification, which means that we insert the mean value of the baseline into the record of those without baseline entry. Table J4 shows the balancing test result without inserting these values, which is essentially the DID sample. The result is quantitatively similar.

[20] As a robustness check, we show the baseline balance for the sample that has both baseline and end-line records of each variable in table D1. It should be noted that the number of observations is smaller than the sample size in table 1 because of attrition. In addition to attrition due to some missing variables in the record, we drop observations with the baseline DT records of some treatment-school students (five schools) because they were offered an inappropriately easier DT. Furthermore, we drop observations if there are any missing responses in survey questions that compose the CPCS or RSES indexes.

TABLE 1
BASELINE BALANCE

Dependent Variable	Treatment	Control	Difference	Observations
DT score[a]	47.092	47.275	−.184	811
	[12.797]	[16.402]	(2.562)	
DT time[a]	9.899	9.960	−.061	811
	[.753]	[.292]	(.054)	
DT score per minute[a]	4.835	4.756	.079	811
	[1.595]	[1.678]	(.274)	
PTSII-C score[b]	34.815	38.940	−4.124	837
	[10.191]	[15.195]	(3.489)	
RSES index[c]	20.997	20.878	.120	832
	[2.506]	[2.731]	(.371)	
CPCS index[c]	27.700	27.004	.696	832
	[2.876]	[3.217]	(.391)	
Female	.599	.629	−.030	843
	[.491]	[.484]	(.030)	
Age	9.897	9.938	−.042	839
	[1.108]	[1.193]	(.304)	
Age^2	99.166	100.186	−1.020	839
	[22.387]	[24.329]	(6.062)	

Note. Sample consists of those who have at least the end-line data. We replace the DT results of those who took a wrong DT with mean DT scores. Standard deviations are shown in brackets. Asymptotic standard errors based on testing the hypotheses that differences between the treatment and control are zero are shown in parentheses and clustered at the school level.

[a] DT is the math diagnostic test. We use three outcomes of DT for measuring cognitive abilities: DT score, DT time, and DT score per minute.

[b] PTSII-C score is the math proficiency test score.

[c] PTSII-C is based on 27 survey questions, of which 10 are consistent with the CPCS index and 8 with the RSES index. For each noncognitive-type question, see app. C.

sample size discussed above because of attrition. In addition to attrition due to some missing values in the record, we consider the baseline DT records of some treatment-school students (five schools) as missing because they were offered inappropriately easier DTs. This is one of the main limitations of this study. Therefore, to maximize the sample size, we adopt the ANCOVA specification in the main analysis (discussed in the next subsection). We do so by replacing missing values with the mean of all nonmissing baseline outcomes. In the estimation, we include a dummy variable that indicates the baseline outcome is missing. Regarding noncognitive variables (i.e., RSES and CPCS indexes), some students could not answer any of the corresponding survey questions to construct the indexes because of time constraints. Therefore, we drop such cases from the analysis. This leads to a slightly smaller sample size than that of the PTSII-C score. Consequently, the number of observations in the final sample is 811 for the DT, 837 for the PTSII-C test, and 832 for the RSES and CPCS indexes. Table 1 suggests that, even with the sample along with the ANCOVA specification, randomization is successful. We check the robustness of our findings

using the full sample and present the results in tables J3 and J4, which are qualitatively similar.

F. Sample Attrition

While some attrition emerges in our sample at the end line, the attrition rate is not significantly different between the treatment and control groups (table D3). The sample is all the observations, including those missing baseline outcomes replaced with the mean values, to be consistent with the working sample in ANCOVA. The dependent variable is a dummy that takes the value of 1 if the student has the end-line outcome and the value of 0 if not.[21] Table D3 shows that attrition does not systematically occur with respect to the treatment status or cause selection issues in our estimates.

III. Empirical Specification and Results

In the main analysis, we adopt the ANCOVA specification (McKenzie 2012; Ma et al. 2024) in addition to the simple end-line comparison. Let t denote the time period, where $t = 0$ illustrates the baseline and $t = 1$ represents the end line. Let Y_{it} be a measure of cognitive or noncognitive abilities of student i at time t; d_i the treatment status (taking 1 for students in the treatment group and 0 in the control group); m_i a missing dummy (taking 1 if missing in Y_{i0} and 0 otherwise); and ε_{it} and ϵ_{it} the error terms. If Y_{i0} is missing, we insert the mean value of the baseline into it. Then, the simple end-line comparison is based on

$$Y_{i1} = \alpha + \delta^{\text{end line}} d_i + \varepsilon_{i1}, \tag{1}$$

while the ANCOVA specification can be written as

$$Y_{i1} = \beta + \delta^{\text{ANCOVA}} d_i + \gamma Y_{i0} + \theta m_i + \epsilon_{i1}. \tag{2}$$

Here, the average treatment effects on the treated can be captured by the estimated δ.[22] We use cluster-robust standard errors at the school level. However,

[21] Further robustness checks on sample attrition, including those of our previous working paper (Sawada et al. 2020) adopting DID specifications, are included in apps. D and J.

[22] In the previous working paper (Sawada et al. 2020), we employ the canonical difference-in-differences (DID) model to estimate the effect of the Kumon intervention on the measures of cognitive and noncognitive abilities of student i at time t, Y_{it}: $Y_{it} = \alpha_0 + \alpha_1 T_t + \phi d_i + \delta^{\text{DID}} T_t \times d_i + u_i + e_{it}$. Here, T_t is a time dummy taking 1 for end line and 0 for baseline, u_i is the student fixed effects, and e_{it} is an error term. The average treatment effects on the treated can be captured by the estimated δ. For the estimation, we take the first difference of the original-level equation, whereby the dependent variable captures improvements in cognitive or noncognitive outcomes:

$$\Delta Y_{it} = \alpha_1 + \delta^{\text{DID}} d_i + \Delta e_{it}, \tag{3}$$

where Δ is a first-difference operator. The results based on this specification are presented in app. E. The results based on DID are qualitatively similar to those based on ANCOVA.

given the relatively smaller number of clusters, we use a wild cluster bootstrap procedure, following Cameron, Gelbach, and Miller (2008).[23]

To investigate heterogeneous treatment effects, we estimate equation (2) for four different subsamples: (i) students who have high initial cognitive ability and high initial noncognitive ability (high-high type); (ii) students who have high initial cognitive ability and low initial noncognitive ability (high-low type); (iii) students who have low initial cognitive ability and high initial noncognitive ability (low-high type); and (iv) students who have low initial cognitive ability and low initial noncognitive ability (low-low type). The cutoff points for high and low are the median values of the respective outcome measures at the baseline.[24] The parameters of interest are δ for different initial ability types.

A. Effects on Cognitive and Noncognitive Abilities

In this subsection, we present the main result based on the empirical specification discussed above. Table 2 reports the treatment effects of Kumon. Panel A presents the results from end-line comparison based on equation (1). Conversely, panel B confirms these findings in panel A by using the ANCOVA specification based on equation (2). It should be noted that all the outcome variables are standardized so the magnitudes of the effects are reported in their standard deviations.[25]

Columns 1–4 of table 2 show the ANCOVA results on cognitive outcomes. As shown in column 1 in panel A, we find significant improvements in the cognitive outcomes measured by DT score, which is as much as 0.429 SD. Effect size based on the ANCOVA specification is similar (0.465 SD) as illustrated in panel B. Furthermore, as discussed above, time reduction in solving questions is the other important dimension in development of cognitive abilities. Therefore, we examine the treatment effects using the measures that consider the time-reduction aspect: DT score per minute and PTSII-C score per minute. The former is the DT score divided by time spent for students to solve the

[23] Unlike the standard cluster-robust standard errors, which are downward biased, this approach reduces overrejection of the null hypothesis through asymptotic refinement without requiring that all cluster data be balanced and the regression error vector be independent and identically distributed (i.i.d.; Cameron, Gelbach, and Miller 2008).

[24] We use different cognitive measures to divide the observations. We use the DT score per minute as the measure of cognitive abilities to specify the median when DT score per minute, DT score, and DT time are the outcome variables, while we use PTSII-C when PTSII-C and noncognitive abilities are the dependent variables.

[25] We report two types of *p*-values in table 2. First, we calculate *p*-value (individual hypothesis testing) by running each regression separately with school-level clustering. Next, *p*-value (individual hypothesis testing, wild bootstrap) is calculated by running each regression separately with school-level clustering by using wild bootstrap.

TABLE 2
EFFECT OF KUMON ON STUDENTS' LEARNING OUTCOMES

	DT Score[a] (1)	DT Time[a] (2)	DT Score per Minute[a] (3)	PTSII-C Score[b] (4)	RSES Index[c] (5)	CPCS Index[c] (6)
	A. End-Line Estimates					
Treatment	.429***	−2.461***	2.283***	.900***	.086	.176
	(.128)	(.426)	(.406)	(.208)	(.150)	(.145)
Constant	.610***	−.733***	.847***	.859***	−.052	−.094
	(.106)	(.228)	(.143)	(.126)	(.085)	(.084)
Observations	811	811	811	837	832	832
R^2	.080	.267	.204	.147	.002	.007
p-value (individual hypothesis testing)	.002	.000	.000	.000	.571	.232
p-value (individual hypothesis testing, wild bootstrap)	.000	.002	.000	.000	.579	.242
	B. ANCOVA Estimates					
Treatment	.465***	−2.209***	2.085***	.999***	.056	.131
	(.144)	(.527)	(.528)	(.210)	(.139)	(.129)
Baseline outcome	.135**	.046	.295***	.335***	.107*	.101**
	(.049)	(.123)	(.095)	(.083)	(.049)	(.040)
Constant	.601***	−.725***	.837***	.810***	.026	−.003
	(.104)	(.220)	(.133)	(.115)	(.089)	(.084)
Observations	811	811	811	837	832	832
R^2	.106	.281	.221	.228	.027	.034
p-value (individual hypothesis testing)	.003	.000	.000	.000	.687	.314
p-value (individual hypothesis testing, wild bootstrap)	.002	.002	.000	.000	.673	.334

Note. Sample is the same as that in table 1. Panel A presents the results from the end-line estimate based on eq. (1), while panel B presents the results from the ANCOVA specification, which is based on eq. (2). Asymptotic standard errors based on testing the hypotheses that the differences between treatment and control are zero are presented in parentheses and clustered at the school level.

[a] DT is the math diagnostic test. We use three outcomes of DT for measuring cognitive abilities: DT score, DT time, and DT score per minute.

[b] PTSII-C score is the math proficiency test score.

[c] PTSII-C is based on 27 survey questions, of which 10 are consistent with the CPCS index and 8 with the RSES index. For each noncognitive-type question, see app. C.

* Statistical significance obtained by clustered wild bootstrap-*t* procedures at the 10% level.

** Statistical significance obtained by clustered wild bootstrap-*t* procedures at the 5% level.

*** Statistical significance obtained by clustered wild bootstrap-*t* procedures at the 1% level.

DT. The latter is the test score of the PTSII-C, which has 228 questions, which is beyond the number of questions that students can deal with within the given time limit so that basically they could not finish all of them. Therefore, to have a high PTSII-C score, students must be accurate and quick in solving questions. The latter requirement enables us to measure time efficiency.[26] Before examining

[26] Furthermore, contrary to standard exams including DT, the students' scores on PTSII-C will not reach the full score. Therefore, this measurement partially avoids the typical censoring problem in estimating treatment effects.

these results, the time-reduction effects are worth examination. As shown in column 2, we see the large negative significant effects of Kumon on DT time, which suggests that Kumon is effective in developing cognitive abilities by enabling students to solve questions in a more time-efficient way. Given this, the results shown in columns 3 and 4 are more suggestive. Kumon improves children's abilities in both accuracy and time efficiency. The magnitude of the effect is sizable: treatment effect measured by DT score per minute using equation (2) is 2.085, as shown in column 3 in panel B. While this effect size may seem surprisingly high compared with the effect size of education interventions elsewhere, the effect size on DT score per minute is due to a substantial reduction in test completion time measured as DT time (−2.209 SD), as discussed above. Similarly, the treatment effects measured by PTSII-C using equation (2) is 0.999, as shown in column 4 in panel B, partly reflecting the time-reduction effects. It should be noted that the effect size of the DT score (0.465 SD), that is, improvement in the raw test score, is consistent with previous findings in the literature wherein similar education intervention was found to be effective in improving learning outcomes. Unlike previous studies that employ test scores to determine cognitive ability, we use test scores per minute (DT score per minute) because our intervention is designed to increase students' abilities to solve math problems in a time-efficient manner, an important ability in pursuing more complex materials in higher education. By contrast, regarding the noncognitive outcomes reported in columns 5 and 6 in panel B of table 2, the homogeneous treatment effect size estimates are insignificant.[27]

The heterogeneous treatment effects are reported in panels A–D of table 3. We find positive and significant coefficients of cognitive outcomes for all four initial ability types. Magnitudes with the measure of DT score per minute are larger for students who have high initial cognitive abilities (high-high type and high-low type). However, they are smallest for students who have low initial abilities in both measures (low-low type). Regarding noncognitive outcomes, however, we

[27] As a robustness check, we report the results focusing on (i) the students without wrong DT distribution and (ii) the student sample with records of all test results in baseline and end line. As shown in tables E1 and E2 together with the baseline balancing results on tables J3 and J4, the effect estimates are qualitatively the same. We report two types of *p*-values in table E1 and three in table E2. First, we calculate the *p*-value (individual hypothesis testing) by running each regression separately with school-level clustering. Next, *p*-value (individual hypothesis testing, wild bootstrap) is calculated by running each regression separately with school-level clustering by using the wild bootstrap. Finally, in table E2, the *p*-value (Romano-Wolf step-down *p*-value) is reported on the basis of multiple hypothesis testing with school-level clustering. While several hypotheses are tested simultaneously, the results are qualitatively the same even when we correct for multiple hypothesis testing using the Romano-Wolf procedure (Romano and Wolf 2005).

find suggestive evidence of the catch-up effect: students who have initially low cognitive and noncognitive abilities (low-low type) show a positive and significant treatment effect on the change in noncognitive scores (RSES index and CPCS index). Conversely, others do not show significant effects in noncognitive scores.

These results support a "building block" story of noncognitive ability. Regardless of the initial cognitive ability, all students have to concentrate for 30 minutes daily during Kumon sessions. This would help build up noncognitive ability even among students who are initially lagging in cognitive ability. In this way, the Kumon intervention first improves the noncognitive ability of those initially lagging in both cognitive and noncognitive abilities (i.e., catch-up on noncognitive ability for low-low type). In turn, this improves the cognitive ability of students who have sufficiently improved noncognitive ability (i.e., higher effects on cognitive ability compared with those of low-high type to low-low type students).

B. Long-Term Effect

To assess the long-term effect of the intervention, we use additional information from a national examination that certifies the completion of primary education (i.e., the PSC exam) 8 and 20 months after the intervention for grade 4 and grade 3 students in our study, respectively. Specifically, we use information about PSC exam take-up and dropouts and math scores obtained by students in our sample.[28]

First, we find that the PSC take-up rate is higher among students in treated schools (50.5%) than among students in control schools (47.7%), albeit the difference is statistically insignificant as shown in table D4.[29]

Considering that only about half of students took the PSC exam, we need to carefully avoid potential selection bias when comparing improvements in cognitive ability. Indeed, among those who took the PSC exam, the average initial DT score and DT completion time of the treatment-school students are significantly lower than those of the control-school students (table D5). We show the

[28] We collected students' PSC exam registration IDs from BPS branch offices and teachers at the schools. We then obtained their PSC exam results from government websites on the basis of IDs. We also collected information from schools about dropouts from the PSC exam (nontakers).

[29] The primary reason for not taking the primary terminal examination was family relocation (79%). Conversely, other reasons included dropouts because of labor market participation (8.5%), school change (7.3%), early marriage (1.5%), sickness (0.75%), death (0.24%), and miscellaneous (2.7%). The registration process for this national examination (usually held at the end of November each year) begins much earlier in the year and closes in September (Nath et al. 2015). This means that when a child's family relocates from the area during this period, they will most likely fail to register a child for the examination at another BPS. However, we could not track the students' families to gather more information on this issue or related reasons behind dropouts. Only letter grades are available for math.

TABLE 3

HETEROGENEOUS EFFECT OF KUMON ON STUDENTS' LEARNING OUTCOMES

	Initial RSES Index					Initial CPCS Index				
	DT Score[a] (1)	DT Time[a] (2)	DT Score per Minute[a] (3)	PTSII-C Score[b] (4)	RSES Index[c] (5)	DT Score[a] (6)	DT Time[a] (7)	DT Score per Minute[a] (8)	PTSII-C Score[b] (9)	CPCS Index[c] (10)
	A. High Initial Cognitive and High Initial Noncognitive Group[d]									
Treatment	.326*	−3.477***	2.962***	1.123***	−.031	.303*	−3.236***	2.913***	1.042***	.193
	(.173)	(.610)	(.692)	(.276)	(.193)	(.149)	(.675)	(.692)	(.319)	(.184)
Baseline outcome	.183	−.114	−.048	.198*	−.084	.103	.076	.146	.144	−.072
	(.145)	(.398)	(.383)	(.111)	(.114)	(.195)	(.521)	(.624)	(.148)	(.127)
Constant	.642***	−.445	.904***	.847***	.276*	.738***	−.731**	.938*	.914***	.231
	(.139)	(.263)	(.237)	(.165)	(.161)	(.189)	(.336)	(.458)	(.224)	(.169)
Observations	159	159	159	189	189	149	149	149	188	188
R^2	.067	.402	.296	.245	.004	.059	.342	.266	.202	.012
	B. High Initial Cognitive and Low Initial Noncognitive Group[d]									
Treatment	.247	−3.459***	2.885***	1.135***	.136	.264	−3.614***	2.741***	1.251***	.004
	(.160)	(.710)	(.902)	(.334)	(.259)	(.202)	(.704)	(.902)	(.265)	(.219)
Baseline outcome	.184	1.403	1.642***	.222	.349***	.276	.031	.381	.279***	.223**
	(.182)	(.909)	(.548)	(.138)	(.111)	(.206)	(.439)	(.510)	(.090)	(.103)
Constant	.705***	−.839**	.047	.849***	.280	.598***	−.404*	.636*	.781***	.183
	(.188)	(.363)	(.401)	(.248)	(.218)	(.173)	(.210)	(.337)	(.178)	(.194)
Observations	124	124	124	180	180	134	134	134	181	181
R^2	.041	.462	.316	.187	.056	.051	.478	.310	.231	.022

	C. Low Initial Cognitive and High Initial Noncognitive Group[d]									
Treatment	.607**	−2.301***	2.457**	1.146***	.108	.506*	−2.071**	2.225**	.980***	.095
	(.273)	(.807)	(1.023)	(.303)	(.214)	(.284)	(.818)	(1.023)	(.322)	(.256)
Baseline outcome	.343***	−.428**	.647*	.545**	.181	.210	−.510***	.592*	.611**	.102
	(.120)	(.182)	(.371)	(.211)	(.190)	(.145)	(.159)	(.340)	(.284)	(.242)
Constant	.663***	−.416	.928***	.784***	−.171	.658***	−.441	.999***	.959***	−.054
	(.234)	(.292)	(.304)	(.233)	(.187)	(.212)	(.286)	(.281)	(.301)	(.265)
Observations	118	118	118	150	150	109	109	109	150	150
R^2	.164	.199	.164	.244	.015	.102	.162	.127	.194	.004
	D. Low Initial Cognitive and Low Initial Noncognitive Group[d]									
Treatment	.540***	−1.410***	1.172***	.980***	.306*	.620***	−1.599***	1.406***	1.110***	.345*
	(.170)	(.498)	(.333)	(.219)	(.160)	(.158)	(.476)	(.333)	(.234)	(.184)
Baseline outcome	.113	1.203	.019	.870***	.036	.162**	1.550	−.019	.772***	−.137
	(.073)	(1.151)	(.217)	(.248)	(.122)	(.072)	(1.136)	(.240)	(.225)	(.098)
Constant	.705***	−.908**	.973***	1.195***	−.228*	.688***	−.942**	.852***	1.036***	−.407***
	(.127)	(.370)	(.232)	(.187)	(.116)	(.130)	(.419)	(.207)	(.160)	(.109)
Observations	143	143	143	177	177	152	152	152	177	177
R^2	.174	.132	.109	.290	.021	.217	.171	.150	.315	.036

Note. Sample is the same as in table 1. The estimation is based on the ANCOVA specification, which is illustrated in eq. (2). Asymptotic standard errors based on testing the hypotheses that the differences between treatment and control are zero are indicated in parentheses and clustered at the school level.

[a] DT is the math diagnostic test. We use three outcomes of DT for measuring cognitive abilities: DT score, DT time, and DT score per minute.

[b] PTSII-C score is the math proficiency test score.

[c] PTSII-C is based on 27 survey questions, of which 10 are consistent with the CPCS index and 8 with the RSES index. For each noncognitive-type question, see app. C.

[d] Initial cognitive score stands for the baseline DT score for cols. 1–3 and cols. 6–8 and for the baseline PTSII-C score for cols. 4, 5, 9, and 10. Initial noncognitive score stands for the baseline RSES index for cols. 1–5. The baseline CPCS index is used in cols. 6–10.

* Statistical significance obtained by clustered wild bootstrap-*t* procedures at the 10% level.

** Statistical significance obtained by clustered wild bootstrap-*t* procedures at the 5% level.

*** Statistical significance obtained by clustered wild bootstrap-*t* procedures at the 1% level.

distribution of the PSC math letter grades of both treatment-group and control-group students in figure G1. While treatment-school students are doing better on the middle range (more Bs and Cs), more control-school students are scoring A+. This also indicates a selection issue in terms of PSC exam participation. In figure G2, we show the distribution of the baseline PTSII-C scores among PSC takers and observe that more high-ability students at control schools take the PSC exam. These results suggest that, among the students with initially low cognitive abilities, treatment students are more likely than control students to stay to take the PSC exam. This suggests the Kumon program might have helped build up grit strength and encouraged students to take the exam after graduating from BPS. However, these discrepancies indicate the presence of an endogenous sample-selection problem with respect to PSC exam participation that needs to be addressed in our analysis. Accordingly, we employ quasi-experimental methods. To eliminate selection bias arising from the unobserved time-invariant heterogeneity, we employ four estimation models: DID, propensity score matching (PSM), inverse probability weighting (IPW), and Lee's bounds methods.

To assess the long-term effect of the intervention, we employ the PSC math and PTSII-C test scores as the end-line and baseline outcomes in a standardized form, respectively. First, apart from the sample-selection problem, we undertake the standard DID analysis using the difference between standardized PSC math score (end line) and standardized baseline PTSII-C test score (baseline) as our dependent variable, controlling for individual fixed effects. Estimated treatment effect is positive but statistically insignificant (panel A of table 4).[30] Second, to mitigate potential selection bias arising from the endogenous decision of taking the PSC exam, we also employ PSM and IPW methods, in which we match the sample on the basis of pretreatment student characteristics (i.e., student age, age squared, and a gender dummy).[31] As shown in panel B of table 4, results suggest that students from treatment schools received statistically significantly higher scores than those of students from control schools wherein point estimates of treatment effects range from 0.226 SD to 0.244 SD.[32] Comparison of the OLS estimation results in panel A with the results from the PSM method and IPW regression in panel B of table 4 suggests that endogenous selection in

[30] To construct this long-term effect measure, we must compare the baseline PTSII-C scores with the PSC exam results, as the students took up the PSC exam only at the end of primary education. Therefore, we standardize both the PTSII-C scores and PSC exam results and then take the differences between them as a measure of changes in cognitive ability. However, this could be a potential limitation of our analysis.

[31] We conduct the balancing check for the matched sample based on PSM. As in table D6, we can see the success in the baseline balancing that supports the validity of PSC analyses even after a large attrition.

[32] See n. 17 for the PSC exam grading scale.

TABLE 4
LONG-TERM EFFECT OF KUMON ON STUDENTS' LEARNING OUTCOMES

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	A. OLS Results: First Difference of PSC Math Score and Baseline PTSII-C Score							
Treatment	.179	.178	.248					
	(.366)	(.367)	(.340)					
Controlling for gender		X	X					
Controlling for age and age^2			X					
Observations	445	445	445					
	B. Point Estimation with Selection and Lee's Bounds							
	PSM		IPW Regression		Lee's Bounds			
	ATT Estimates	ATE Estimates	ATT Estimates	ATE Estimates	Lower-Bound Estimate		Upper-Bound Estimate	
Treatment	.226*	.235*	.244**	.233**	.040	.051	.320*	.314*
	(.123)	(.120)	(.121)	(.118)	(.158)	(.162)	(.164)	(.163)
Constant			−.270***	−.240***				
			(.095)	(.090)				
Control in selection equation (gender)						X		X
Observations	445	445	445	445				
No. of selected observations[a]					445	445	445	445
No. of total observations[b]					905	905	905	905

Note. Panel A presents the result from the ordinary least squares (OLS) specification, while panel B considers the sample selection. Asymptotic standard errors based on testing the hypotheses that the differences between treatment and control are zero are presented in parentheses. In the analysis, we replace missing values for age with 0 but indicate a missing dummy in the estimation. ATT = average treatment effect on the treated; ATE = average treatment effect.

[a] Number of observations whose record of the end-line outcome is observable.

[b] Number of total observations, including those without the record of the end-line outcome.

* Statistical significance obtained by clustered wild bootstrap-*t* procedures at the 10% level.

** Statistical significance obtained by clustered wild bootstrap-*t* procedures at the 5% level.

*** Statistical significance obtained by clustered wild bootstrap-*t* procedures at the 1% level.

taking the PSC exam might have generated downward bias in estimating treatment effects.[33] Third, we estimate Lee's bounds (Lee 2009) and consider nonrandom sample selection in taking the PSC exam following the monotonicity assumption, that is, no heterogeneous effect of the treatment on selection. As shown in panel B of table 4, upper-bound estimates are statistically significant. Point estimates of PSM and IPW (panel B) are within these Lee's bounds (0.040 SD to 0.320 SD). Overall, we find a modest but positive long-term effect of the intervention on cognitive ability measured by math test score. Moreover, we show the heterogeneous treatment effects by the baseline PTSII-C score

[33] Recall that low-skilled students took the PSC exam more in the treatment group, which seems to drive this downward bias.

(fig. H1):[34] treatment effects seem higher for students whose baseline PTSII-C scores are in the 40th–60th and 80th–100th percentiles.[35] Although estimation is imprecise, most students benefited from the Kumon intervention.[36]

To better understand the path of the long-term effects, we investigate heterogeneity in terms of the cohort. We have two cohorts—and the timing of the PSC exam is different—in addition to the several age cohorts in our analysis. Therefore, we conduct the heterogeneity analyses in terms of age and initial ability. Figure H2 and table H1 show the results. The estimation results suggest that the treatment effects are higher and positive when the intervention occurs when students are young. Conversely, the effects gradually fade and become negative when it occurs when they are older. This pattern is consistent with the literature on education intervention that a childhood intervention should be conducted as early as possible (Heckman 2006; Hendren and Sprung-Keyser 2020), and the effects may deteriorate when it is late (Chetty, Hendren, and Katz 2016). Furthermore, we examine heterogeneity in terms of initial grade. As in figure H1, the better the initial ability, the larger the long-term effects.

C. *Teacher Assessment Ability*

In addition to student outcomes, we examine the effect of the intervention on teachers' abilities to assess their students' performance. We hypothesize that teachers may be able to improve their own understanding and assessment of students' abilities because intervention will allow them to gain more information about students' abilities from record books. Using absolute distance between teachers' assessment scores and students' test scores (for each student) as a dependent variable, we conduct a DID analysis. As shown in table 5, we find significant improvement in teachers' abilities to assess students' performance in both DT and PTSII-C scores (i.e., a negative sign indicates that the assessment scale is closer to actual test-score scale).

These effects on BPS teachers are unintended but unsurprising, given the nature of the intervention. BPS teachers interact with the program to the extent that they ensure students comply with the intervention (i.e., study at the right level). BPS teachers obtain a partial signal of each student's ability from the level of the

[34] We appreciate an anonymous referee for suggesting this approach.

[35] The average baseline PTSII-C score is not statistically different between the treatment-school and control-school students who took the PSC exam. The standardized mean of PTSII-C scores among PSC takers at the treatment schools (−0.105) and at the control schools (0.334) differs by 0.439 (p-value: .110). However, the standardized mean of the DT score of PSC takers from the treatment schools (−0.021) and the control school (0.266) significantly differs by 0.287 (p-value: .088).

[36] It shows a negative point estimate for the lowest group but is not statistically significant. However, we may need additional care if we introduce this to very low-skilled students.

TABLE 5
ASSOCIATION BETWEEN TEACHER'S ASSESSMENT AND STUDENT PERFORMANCE

	Absolute Difference between Teacher's Perception and Student's Score	
	DT Score[a] (1)	PTSII-C Score[b] (2)
Treatment × end line	−.919**	−.350**
	(.265)	(.132)
Treatment	−.045	−.219
	(.294)	(.142)
End line	−.248	.148*
	(.192)	(.077)
Constant[c]	2.346***	1.535***
	(.241)	(.110)
Observations	990	1,416
R^2	.101	.047

Note. Dependent variable is the absolute difference between the teacher's subjective evaluation and the student's objective performance. Asymptotic standard errors are shown in parentheses and clustered at the school level.
[a] DT score is the math diagnostic test.
[b] PTSII-C score is the math proficiency test score.
[c] Significance level of the coefficients is based on the standard *p*-value.
* Statistical significance obtained by clustered wild bootstrap-*t* procedures at the 10% level.
** Statistical significance obtained by clustered wild bootstrap-*t* procedures at the 5% level.
*** Statistical significance obtained by clustered wild bootstrap-*t* procedures at the 1% level.

worksheets and the speed in solving them. While this may suggest that teachers could have modified teaching in program schools, we find no significant difference in teaching hours or home workloads between treatment and control schools. We agree that better information about students' progress gives teachers in treatment schools the ability to more accurately assess students' abilities. The Kumon learning approach has good potential for reducing teachers' stereotyping of students by providing teachers with better information about their students.[37]

IV. Comparing Benefits and Costs

Following Duflo (2001) and Heckman et al. (2010), we calculate the benefit-cost ratio and the internal rate of return (IRR). Regarding benefits, we use our long-term effect estimate on math PSC scores (table 4) and estimated wage returns to numeracy skills from Nordman, Sarr, and Sharma (2015), who use matched employer-employee data. Benefit per student is calculated as a product

[37] These results provide important insights about teachers' roles and effectiveness in learning, because teachers in many countries have a fixed mindset about the learning potential of low-performing students (Sabarwal, Abu-Jawdeh, and Kapoor 2022).

of the effect of Kumon on math ability (SD), wage returns on numeracy skills (SD), and average annual earnings.[38] We assume that the benefit will last from 1 to 44 years, considering the working age as 16–59 years and an annual discount rate of 5%, following Duflo (2001). We do not use the deadweight loss factor, because this program did not involve tax spending or revenue.

As the minimum cost, we consider worksheet printing costs on the basis of number of worksheets actually used and costs related to transportation, purchasing of clocks, salary for personnel, and training. For the maximum cost calculation, we add 50% higher worksheet printing costs if some students completed a higher level, regardless of use. According to the project budget record, the minimum (maximum) cost per student is 8,786 (9,619) Bangladesh taka, or 113 (124) USD, for 8 months.

Under the minimum (maximum) cost assumption, the benefit-cost ratio exceeds 1 when benefits last for more than 14 (more than 16) years, as shown in figure I1 (fig. I2). However, it should be noted that the wage returns to numeracy skills are estimated on the basis of full-time formal-sector jobs, which is a growing sector but not necessarily a representative type of employment in Bangladesh. The IRR is calculated so that the present values of benefits and costs equalize over a specified time horizon, varying from 1 to 44 years. The IRR becomes positive when workers continue working with benefits for more than 10 (11) years with the minimum (maximum) cost (figs. I1 and I2).

V. Discussion

In this study, we investigate the effectiveness of a novel individualized self-learning method in overcoming the issue of low-quality learning in a developing country context. The intervention consists of supplementary learning materials beyond the regular curriculum. Specifically, we conducted a field experiment in Bangladesh to test the effectiveness of the Kumon method of learning in improving primary school students' cognitive and noncognitive abilities. As an effective program for strengthening student abilities, Kumon is based on a just-right level of study that provides a suitable amount of mental stimulus to enhance academic and self-learning outcomes. Our intervention included a 30-minute Kumon session before regular school hours—6 days a week for

[38] The first estimate is taken from our results on the PSC exam, and we use the most conservative number (PSM-ATT estimates), 0.226, in table 4. Wage returns to numeracy skills, 0.037, are taken from the paper by Nordman, Sarr, and Sharma (2015), their table 3, col. 8. Average annual earnings are calculated on the basis of average hourly wage in the paper by Nordman, Sarr, and Sharma (2015), their table 2 (50.91), multiplied by 40 hours per week and 52 weeks. We then calculate the life-cycle profile of earnings on the basis of estimates of the returns to tenure and tenure squared (0.037 and −0.00067) in the regression of Nordman, Sarr, and Sharma (2015).

8 months. This was offered among BPS students, who come from disadvantaged backgrounds.

We find significant and robust improvements in students' cognitive abilities. Given that our intervention was designed to increase students' math problem-solving skills in a time-efficient manner, we demonstrate the effect using time-adjusted test scores, whereby effect comes through both test-score gains and reduction in problem-solving speed. When using such unconventional measurements, we observe a relatively large effect size compared with education interventions elsewhere. One may argue that our cognitive ability results could be attributed to additional math learning per se over 8 months rather than to self-learning at the right level. Our back-of-the-envelope calculations under the conservative assumption of constant returns to scale suggest at least 16% (0.19 SD) of the effect can be attributed to the effects of self-learning at the right level, meaning that we see a positive effect of individualized self-learning methods. However, this is based on two relatively strong assumptions of parallel trend and linear-in-time cognitive growth.[39] Moreover, the intervention is particularly designed to improve math problem-solving skills through building endurance and perseverance. Hence, we find there are catch-up effects on noncognitive abilities among students with initially low cognitive and noncognitive abilities. In terms of achieving the standard sought by the national curriculum, which is evaluated by the nationally administered primary school certificate examination, we observe that intervention improves students' ability in an expected direction. In particular, our results show some long-term effect of the intervention when comparing students' achievements on the national-level examination taken 8 and 20 months after the intervention with their baseline math proficiency test scores. Finally, although the BPS teacher's role during a Kumon session was limited to monitoring and mechanically determining the level of worksheets on the basis of the predefined procedure, we find positive effects on BPS teachers' capacity to assess student performance. This finding implies that BPS teachers might have benefited from Kumon intervention by gaining more objective information about students' skills. However, we have no evidence suggesting that the intervention affected their regular teaching practice. Future research should focus on teachers' perceptions and teaching practice.

This paper-and-pencil-based self-learning program is well suited for settings constrained by inadequate ICT infrastructure and therefore is easily scalable in

[39] (i) Parallel trend assumption: we assume that the counterfactual growth of the treatment group's cognitive ability is the same as that of the control group. (ii) Linear effect assumption: we are assuming that the growth of cognitive abilities over 8 months, without treatment, is linear in time (instead of diminishing ability growth as the time goes by). See app. E and tables E2 and E3 for the details of the calculation process.

developing countries. Hence, the results obtained are generalizable in the case of other intended beneficiaries in a similar setting. While we follow standard Kumon worksheets, protocol, and routine procedures, the deployed resources were generally limited compared with those of commercially operated Kumon centers elsewhere. Therefore, we believe that Kumon could be a cost-effective complementary intervention to existing lecture-style primary education curricula.

We note potential limitations of our analysis. First, some observations are dropped because of noncompliance and attrition resulting in a smaller sample size than in the initial design. However, as we show in our robustness analysis, these do not substantially affect our main conclusion. Nevertheless, a larger-scale RCT, including rural areas and public schools, may be useful for enhancing the external validity of our results. Second, our long-term effect analysis is based on public examinations on the national curriculum administered after the RCT intervention. This results in substantial attrition in participation in the nationally administered test, as many students could not take the examination from their school because of family relocation issues. However, no significant difference is found in PSC exam take-up rate between treatment-school and control-school students. We address potential selection bias using quasi-experimental analysis. Third, in our benefit-cost analysis, we rely on PSC test scores after 8 and 20 months showing the long-term benefit of the intervention. These effects are already reduced compared with our main results. Therefore, the long-term benefit estimates shown here could diminish over time. Finally, considering its focus, the current study does not detail the mechanisms behind the effect of the Kumon method. In a companion paper, we investigate the peer effects on classroom learning among treatment students (Kawarazaki et al. 2023). Uncovering these mechanisms is a key task for future research.

References

AERA (American Educational Research Association), APA (American Psychological Association), and NCME (National Council on Measurement in Education). 2014. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Afroze, Rifat. 2012. "How Far BRAC Primary Schools Admit Students Following the Set Criteria." Report, BRAC, Dhaka.

Ahmad, Alia, and Iftekharul Haque. 2011. "Economic and Social Analysis of Primary Education in Bangladesh: A Study of BRAC Interventions and Mainstream Schools." Research Report, BRAC, Dhaka.

Asadullah, Mohammad Niaz. 2016. "Do Pro-Poor Schools Reach Out to the Poor? Location Choice of BRAC and ROSC Schools in Bangladesh." *Australian Economic Review* 49, no. 4:432–52.

Asadullah, Mohammad Niaz, and Nazmul Chaudhury. 2013. "Primary Schooling, Student Learning and School Quality in Rural Bangladesh." Working Paper no. 349, Center for Global Development, Washington, DC.

Asim, Salman, Robert S. Chase, Amit Dar, and Achim Schmillen. 2017. "Improving Learning Outcomes in South Asia: Findings from a Decade of Impact Evaluations." *World Bank Research Observer* 32, no. 1:75–106. https://doi.org/10.1093/wbro/lkw006.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton. 2016. "Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of 'Teaching at the Right Level' in India." NBER Working Paper no. 22746 (October), National Bureau of Economic Research, Cambridge, MA.

Banerjee, Abhijit V., Shawn Cole, Esther Duflo, and Leigh Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics* 122, no. 3:1235–64.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *Review of Economics and Statistics* 90, no. 3:414–27.

Chetty, Raj, Nathaniel Hendren, and Lawrence F. Katz. 2016. "The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment." *American Economic Review* 106, no. 4:855–902.

Chowdhury, Ahmed Mushtaque Raza, Andrew Jenkins, and Marziana Mahfuz Nandita. 2014. "Measuring the Effects of Interventions in BRAC, and How This Has Driven 'Development.'" *Journal of Development Effectiveness* 6, no. 4:407–24.

Duflo, Esther. 2001. "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment." *American Economic Review* 91, no. 4:795–813.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking." *American Economic Review* 101, no. 5: 1739–74.

Engelhardt, Lena, and Frank Goldhammer. 2019. "Validating Test Score Interpretations Using Time Information." *Frontiers in Psychology* 10:1131.

Evans, David K., and Anna Popova. 2015. "What Really Works to Improve Learning in Developing Countries?: An Analysis of Divergent Findings in Systematic Reviews." Policy Research Working Paper no. 7203, World Bank, Washington, DC.

Ganimian, Alejandro J., and Richard J. Murnane. 2016. "Improving Education in Developing Countries: Lessons from Rigorous Impact Evaluations." *Review of Educational Research* 86, no. 3:719–55.

Glewwe, Paul, ed. 2014. *Education Policy in Developing Countries*. Chicago: University of Chicago Press.

Harter, Susan. 1979. *Perceived Competence Scale for Children*. Denver, CO: University of Denver.

Heckman, James J. 2006. "Skill Formation and the Economics of Investing in Disadvantaged Children." *Science* 312, no. 5782:1900–1902.

———. 2007. "The Economics, Technology, and Neuroscience of Human Capability Formation." *Proceedings of the National Academy of Sciences of the USA* 104, no. 33:13250–55.

Heckman, James J., John Eric Humphries, and Tim Kautz, eds. 2014. *The Myth of Achievement Tests: The GED and the Role of Character in American Life*. Chicago: University of Chicago Press.

Heckman, James J., Seong Hyeok Moon, Rodrigo Pinto, Peter A. Savelyev, and Adam Yavitz. 2010. "The Rate of Return to the HighScope Perry Preschool Program." *Journal of Public Economics* 94, nos. 1–2:114–28.

Heckman, James J., Jora Stixrud, and Sergio Urzua. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *Journal of Labor Economics* 24, no. 3:411–82.

Hendren, Nathaniel, and Ben Sprung-Keyser. 2020. "A Unified Welfare Analysis of Government Policies." *Quarterly Journal of Economics* 135, no. 3:1209–318.

Kawarazaki, Hikaru, Minhaj Mahmud, Yasuyuki Sawada, and Mai Seki. 2023. "Haste Makes No Waste: Positive Peer Effects of Speed Competition on Classroom Learning." *Oxford Bulletin of Economics and Statistics* 85, no. 4:755–72.

Kremer, Michael, Conner Brannen, and Rachel Glennerster. 2013. "The Challenge of Education and Learning in the Developing World." *Science* 340, no. 6130: 297–300.

Lakshminarayana, Rashmi, Alex Eble, Preetha Bhakta, Chris Frost, Peter Boone, Diana Elbourne, and Vera Mann. 2013. "The Support to Rural India's Public Education System (STRIPES) Trial: A Cluster Randomised Controlled Trial of Supplementary Teaching, Learning Material and Material Support." *PLOS ONE* 8, no. 7:e65775.

Lee, David S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *Review of Economic Studies* 76, no. 3:1071–102.

Ma, Yue, Robert Fairlie, Prashant Loyalka, and Scott Rozelle. 2024. "Isolating the 'Tech' from EdTech: Experimental Evidence on Computer Assisted Learning in China." *Economic Development and Cultural Change* 72, no. 4.

McEwan, Patrick J. 2015. "Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments." *Review of Educational Research* 85, no. 3:353–94.

McKenzie, David. 2012. "Beyond Baseline and Follow-Up: The Case for More T in Experiments." *Journal of Development Economics* 99, no. 2:210–21.

Muralidharan, Karthik, Abhijeet Singh, and Alejandro J. Ganimian. 2019. "Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India." *American Economic Review* 109, no. 4:1426–60.

Nath, Samir Ranjan. 2012. "Competencies Achievement of BRAC School Students: Trends, Comparisons and Predictors." Research Monograph no. 51, BRAC, Dhaka.

Nath, Samir Ranjan, A. Mushtaque R. Chowdhury, Manzoor Ahmed, and Rasheda K. Choudhury, eds. 2015. "Whither Grade V Examination? An Assessment of Primary Education Completion Examination in Bangladesh." Education Watch 2014, Campaign for Popular Education, Bangladesh.

Nordman, Christophe J., Leopold R. Sarr, and Smriti Sharma. 2015. "Cognitive, Noncognitive Skills and Gender Wage Gaps: Evidence from Linked Employer-Employee

Data in Bangladesh." Working Paper no. DT/2015/19, DIAL (Développement, Institutions et Mondialisation), Paris.

Romano, Joseph P., and Michael Wolf. 2005. "Stepwise Multiple Testing as Formalized Data Snooping." *Econometrica* 73, no. 4:1237–82.

Rosenberg, Morris. 1965. *Society and the Adolescent Self-Image*. Princeton, NJ: Princeton University Press.

Sabarwal, Shwetlena, Malek Abu-Jawdeh, and Radhika Kapoor. 2022. "Teacher Beliefs: Why They Matter and What They Are." *World Bank Research Observer* 37, no. 1:73–106.

Sakurai, Shigeo, and Yutaka Matsui, eds. 1992. *Shinri Sokutei Shakudo Shu IV (Psychological measurement scale IV): Jido-you Konnpitensu Shakudo (Competence scale for children) "Jikokachi (Self-Worth),"* 22–27. Tokyo: Saiensu-sha.

Sawada, Yasuyuki, Minhaj Mahmud, Mai Seki, An Le, and Hikaru Kawarazaki. 2020. "Fighting the Learning Crisis in Developing Countries: A Randomized Experiment of Self-Learning at the Right Level." Discussion Paper no. CIRJE-F-1127, Center for International Research on the Japanese Economy, University of Tokyo.

UNESCO. 2013. "Teaching and Learning: Achieving Quality for All." Education for All Global Monitoring Report 2013/4, UNESCO, Paris.

United Nations. 2018. "The Sustainable Development Goals Report 2018." United Nations, New York.

Watkins, Kevin. 2000. "The Oxfam Education Report." Oxfam International, Oxford.

World Bank. 2018. "World Development Report 2018: Realizing the Promise of Education for Development." World Bank, Washington, DC.