Fighting the Learning Crisis in Developing Countries:

A Randomized Experiment of Self-Learning at the Right Level

By

Yasuyuki Sawada (The University of Tokyo and RIETI)
Minhaj Mahmud (Asian Development Bank)
Mai Seki (Ritsumeikan University)
Hikaru Kawarazaki (University College London and the Institute for Fiscal Studies)

August 2022

CREPE DISCUSSION PAPER NO. 128

CENTER FOR RESEARCH AND EDUCATION FOR POLICY EVALUATION (CREPE)
THE UNIVERSITY OF TOKYO
http://www.crepe.e.u-tokyo.ac.jp/

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

Yasuyuki Sawada, Minhaj Mahmud, Mai Seki and Hikaru Kawarazaki

June 2022

Abstract

This study investigates the effectiveness of a globally popular method of self-learning at the right level in improving learning outcomes—the cognitive and noncognitive abilities of disadvantaged students—in a developing country, Bangladesh. Using a randomized controlled trial design, we find substantial improvements in cognitive abilities measured by mathematics test scores and catch-up effects in terms of noncognitive abilities or personality traits measured through a self-esteem scale. Moreover, our study is the first to use alternative cognitive ability measures, that is, time reduction as well as time-adjusted test score, which are critical dimensions of cognitive development. Subsequently, we investigate the long-term effects using students' math results of the national-level exam. We find a reasonable longer-term impact on cognitive abilities 20 months after the intervention for younger students. Our estimates indicate that the program's benefits exceed its costs.

JEL code: I20, O12

Author Affiliations and Acknowledgments

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." Yasuyuki Sawada (corresponding author), Faculty of Economics, the University of Tokyo (sawada@e.u-tokyo.ac.jp); Minhaj Mahmud, the Economic Research and Regional Cooperation Department, the Asian Development Bank (mmahmud@adb.org); Mai Seki, Faculty of Economics, Ritsumeikan

University (maiseki@fc.ritsumei.ac.jp); Hikaru Kawarazaki, Department of Economics, University College London and the Institute for Fiscal Studies (hikaru.kawarazaki.20@ucl.ac.uk). The opinions expressed in this paper are those of the authors' and do not reflect the views of affiliated organizations. We would like to thank An Le, Saori Nishimura, and Kazuma Takakura for superb research assistance. We appreciate the editor, Marcel Fafchamps, an associate editor, and two anonymous referees of the journal for their constructive comments. We would also like to thank Esther Duflo, Pascaline Dupas, Deon Filmer, Dean Karlan, Halsey Rogers, and Paul Romer. Further, we thank the session participants at the American Economic Association Meeting 2018, World Bank Education GP BBL, the 2017 European, North American Summer, Asian Meetings of the Econometric Society, the Australasian Development Economics Workshop 2017, the Midwest International Economic Development Conference 2017, the GRIPS-University of Tokyo Workshop 2017, Hitotsubashi University, Kansai Labor Economics Workshop, and 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 (JICA) for their cooperation in implementing the study.

Disclosure and Data Replication Statement

The study protocol went through full committee review and was approved by the IRB of the University of Tokyo (refs. 15–90). The study has been registered at the American Economic Association's Randomized Controlled Trials (RCTs) registry (AEARCTR-0002925). This work was supported by a grant from Grants-in-Aid for Scientific Research (S) from Japan Society for the Promotion of Science (KAKENHI 26220502). The authors declare that they have no financial interest related to this study. They designed and conducted a randomized controlled trial (RCT) evaluation independent of Kumon Institute of Education Co., Ltd. On a formal request written to Kumon Institute of Education Co., Ltd. through the corresponding author of this paper, we make available all the data files to the journal (without requiring personal information, such as individual names), programs, and details for the computations necessary for replication.

1 Introduction

Learning crisis refers to the global phenomenon wherein over 60 percent of children who complete their primary education in low- and middle-income countries fail to achieve a minimum proficiency in mathematics and reading (World Bank, 2018; UNESCO, 2013). Also, improving the quality of education is a *sine qua non* for achieving SDGs (United Nations, 2018). Owing to their high effectiveness in improving learning outcomes, teaching at the right level (TaRL) programs are gaining increasing 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 can potentially constrain usage of such effective programs.

In this study, we evaluate the effectiveness of an individualized self-learning program, the Kumon method of learning (hereinafter referred to as 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 in Bangladesh. Kumon is a globally popular, non-formal education program that is designed to ensure that each student always studies at the level that is "just right" for them.[2] In Kumon, each student begins at an individually suitable starting point identified through a diagnostic test (DT) 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, non-formal education has been crit-

[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 impact evaluation studies that focus on improving learning outcomes in South Asian countries. Other reviews that focus on the impacts of interventions on learning outcomes include Kremer, Brannen and Glennerster (2013); Ganimian and Murnane (2016); Evans and Popova (2015); McEwan (2015); Glewwe (2014).

[2]As of December 2021, there are 3.7 million Kumon subject enrollments, and the program has been adopted in 61 countries and regions (as of March 2022), according to the Kumon Institute of Education Co., Ltd. See `https://www.kumongroup.com/eng/about/` for the details (Last access: May 30th, 2022).

ical to this process. With this respect, non-governmental 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 enrollment and reducing primary school dropouts, the country's 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 and 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 the same as in other developing countries.[4]

In this context, we adopt and evaluate the impact of Kumon in improving both cognitive and noncognitive abilities of BPS students in Bangladesh, given Kumon's unique setting in providing non-formal 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 the first to experimentally investigate its impact 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 mathematics (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 mathematics materials for each student.[5] Indeed, Kumon's goal has been on improving both cognitive ability and cer-

[3]BPSs are regarded one of the largest and most successful non-formal education programs that are targeted at disadvantaged populations in Bangladesh. They have introduced a seasonally adjusted school calendar, which has been a key to their success (Watkins, 2000; Chowdhury, Jenkins and Nandita, 2014). Section 2 provides more details about BPSs.

[4]For example, Asadullah and Chaudhury (2013) find an imperfect correlation between years of schooling and cognitive outcomes: among the children who completed primary schooling, only 49 percent could provide 75 percent or higher correct answers in a simple arithmetic test, and the likelihood of providing more than 75 percent correct answers was only 9 percent higher than 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

tain noncognitive abilities (e.g., perceived competence, self-confidence, and self-esteem). According to the website of Kumon Institute of Education, "[a]s 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 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.

We measure the cognitive ability improvements by comparing the mathematics test scores obtained by students at the baseline and endline, which is known as the diagnostics test (DT) score. The findings indicate that Kumon substantially improves students' cognitive abilities, and this is measured through the DT score by 0.465 s.d.[7] Given that our intervention was designed to increase students' math problem-solving skills in a time-efficient manner, we show the impact using test scores per minute wherein the impact comes through both test score gains and reductions in the problem-solving speed. Therefore, the magnitude of the impact through test score per minute (2.085 s.d.) is much higher than the effect size of education interventions elsewhere. Interestingly, this is largely owing to a substantial reduction in test completion time as reflected in the effect size of the DT time (−2.209 s.d.). To our knowledge, our study is the first to use time reduction and time-adjusted measures of cognitive ability, which, we believe, are critical dimensions for cognitive development. Additionally, we measure cognitive ability

to 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 non-formal education program in developed countries, its impacts on learning outcomes in a disadvantaged setting in a developing country context is worth evaluating.

[6]See `https://www.kumongroup.com/eng/about-kumon/future/` for details (Last access: May 30th, 2022).

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

using a second mathematics test score, which is known as the proficiency tests of self-learning skills (hereinafter referred to as PTSII-C) score. PTSII-C not only captures the accuracy but also tests how many problems students could attempt to solve within the specified time. That is, their score already reflects both accuracy and speed. Indeed, the effect size in the case of PTSII-C score is comparably high (i.e., 0.999 s.d.). Regarding noncognitive abilities measured through certain personality traits, we find catch-up effects among students with initially lower abilities compared to those of the median.

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

The remainder of this paper is organized as follows. In Section 2, we outline our experimental design, including the setting and the intervention, and then explain the data and baseline test results. Section 3 presents the econometric evaluation framework, followed by the empirical results. Section 4 compares the benefits and costs of this intervention, and finally Section 5 discusses findings and caveats.

[8]Although these two tests are somewhat different, they measure math ability. We could avoid a potential concern in comparing two different tests in baseline and end line as we use standardized scores.

2 Experiment Design, Data, and Balancing Test

2.1 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). 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). Particularly, 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., PSC). This also suggests that BPS provides learners with the same skills that are taught in government schools; that is, teaching to 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 learn to set goals and take up challenges at the next level. Given the unique setting of this non-formal 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.

2.2 Intervention: The Kumon Method of Learning

As a supplementary module in mathematics, the Kumon method of learning has been introduced in selected BPSs among third- and fourth-grade students.

Kumon Method of Learning

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 based on their individual performance in a DT, rather than based on 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 because of the following reasons: to i) ensure that students fully understand the basic concepts and develop a firm foundation for the development of their cognitive abilities and ii) motivate them to continue studying, which also aids the development of their noncognitive abilities (e.g., self-esteem and sense of competence). Kumon worksheets are designed, ranging from simple counting to advanced mathematics, 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.[9] Kumon instructors do not conduct 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.

[9]See example worksheets of Kumon in Appendix A (Figures A1 and A2).

Another feature of Kumon is a system that tracks each student's progress and achievements using 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. Further, 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 allows instructors to obtain more objective information about their students' abilities and understanding of the mathematics involved.

Intervention in BPSs

Our intervention was a pilot program in BPSs to examine the effectiveness of Kumon in a disadvantageous setting encountered by resource constraints and run during regular school hours. Unlike the regular Kumon sessions elsewhere, BPS provided 30-minute Kumon session daily without any homework assignments. The learning materials were supplied by the Kumon Institute of Education Co., Ltd., 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 did not conduct any lecture. Instead, they observe their students' progress on individualized worksheets without intervention in principle. When a student becomes stuck solving problems after many attempts, they adjust the level of worksheets downward to facilitate individual learning based on the pre-fixed procedure they learned during the training sessions.[10] The BPS teachers are not responsible for grading or recording the marks. The designated marking assistants will give the grade and records 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 ends, students

[10]There were short conversations between a BPS teacher and each student, but there was no direct teaching during the Kumon session. Also, the teachers needed to determine students' worksheet levels fairly mechanically based on the scores and time in principle as trained. Therefore, if any, these interactions should not play an important role in students' learning.

either move on to a new worksheet once they had achieved a full score on the previous one or continue to attempt and correct their answers until they achieved a full score within the designated time frame. On rare occasions when students encounter great difficulty with higher-order problem solving tasks beyond their grade level, the BPS teachers might come only to clarify the examples in the worksheet.

2.3 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 hypothesize a minimum detectable effect of 0.40 s.d. on students' cognitive ability. In our context, we referred to the results from some studies of high-impact education interventions that involve teaching at the right level (e.g., Lakshminarayana et al. (2013) (0.75 s.d.) and Duflo, Dupas and Kremer (2011) (0.9 s.d.)) and hypothesize the effect size to be 0.4 s.d. Considering that randomization is conducted at the cluster (school/classroom) level, we assume an intracluster correlation of 0.10 and a statistical significance of less than 0.05 for a two-tail 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 owing to attrition or other factors, we selected a cluster size of 34 to increase the total student sample, with 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 to 17 treatment and 17 control schools. The resulting sample breakdown by class/grade is as follows: 19 (out of 48 schools) for the third grade and 15 (out of 131 schools) for the fourth grade.[11] The schools do not overlap in terms of grade. That is, in a particular school, we only offered the intervention to either grade 3 or grade 4.

The intervention consisted of a 30-minute session on the Kumon method prior to the

[11]Based on 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 third and fourth grades, respectively. One grade 4 school turns out to be third grade school, resulting in odd numbers of schools for each grade.

students' regular lessons. Thus, during the study period, the students in the treatment schools arrived to school earlier than their usual school 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 of Education Co., Ltd. provided an intervention package comprising a mathematics material set and an instructor manual with sheets for the BRAC teachers.[12] The full material set comprises i) mathematics worksheets with questions at various difficulty levels and achievement tests at the end of each level and ii) a record book to keep the students' daily progress. This includes 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 (Figure A3).[13][14]

2.4 Data Description

We constructed cognitive ability measures at both the baseline and endline based on two different mathematics test scores for both the treatment and control school students. These mathematics tests are DT and PTSII-C. The DT measures cognitive (math) abilities, whereby we retain records of both the score and time taken to complete the test.

[12]BRAC field staff has been assigned to assist and follow up on BPS teachers. Prior to launching the program, a three-day-long preparatory training for BPS teachers and field staff was held to familiarize them with the concepts and procedures of the learning method, followed by additional three one-day-long 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.

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

[14]We believe that our intervention is not necessarily idealized but well designed to follow regular channels—classroom setting without ICT infrastructure—and the results obtained are generalizable in the case of other intended beneficiaries in a similar setting. Although we follow the standard Kumon worksheets, protocol, and routine procedures, the deployed resources were generally limited compared to the commercially operated Kumon centers elsewhere.

The DT used for this study is time-specific and requires students to answer 70 questions within a maximum of 10 minutes.[15] 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 their cognitive ability (i.e., PTSII-C).[16] 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 could attempt to solve within a specified time. The second section comprises 27 questions that measure the aspects of noncognitive abilities (see Table C1 of Appendix C). 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) (Sakurai and Matsui, 1992; Harter, 1979). As noncognitive ability measures, we created the RSES and CPCS Indices based on these questions.[17]

To assess the possible long-term impact of the intervention, we also collected students' results from the PSC examination, a nationally administered primary education comple-

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

[16]The PTSII-C includes 348 questions, which comprise 120 extremely simple tasks (Part 1) and 228 simple math questions (Parts 2–6). The former task questions ask 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, they 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 equations (Part 6). Parts 2 and 3 are standard calculation problems that we see in any calculation problem sets. These are also similar to the DT and everyday worksheets. Parts 4 and 5 are unique to the PTSII, and students do not see either in the DT or everyday worksheets. Part 6 is also a unique style of problems not commonly seen in standard calculation problems. However, these types of questions overlap with some parts of everyday worksheets.

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

tion test by the Ministry of Primary and Mass Education.[18] 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).[19]

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

2.5 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 endline records of each variable are reported in Table 1.[20] 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 ran-

[18]Those who wish to pursue further education must pass this exam. Based on the exam results, letter grades from A+ to A, A-, B, C, D, and F are assigned: if the score is in the range of 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 Math and English, in addition to other subjects. Unfortunately, we do not have data on the exact score for an individual subject, but we have data on the letter grades. See `http://www.educationboard.gov.bd/computer/grading_system.php` for details (Last access: May 30th, 2022).

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

[20]The sample is for the 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.

Table 1. Baseline Balance

Dependent Variable	Treatment	Control	Difference	N
DT Score[a]	47.092	47.275	-0.184	811
	[12.797]	[16.402]	(2.562)	
DT Time[a]	9.899	9.960	-0.061	811
	[0.753]	[0.292]	(0.054)	
DT Score per min[a]	4.835	4.756	0.079	811
	[1.595]	[1.678]	(0.274)	
PTSII-C Score[b]	34.815	38.940	-4.124	837
	[10.191]	[15.195]	(3.489)	
RSES Index[c]	20.997	20.878	0.120	832
	[2.506]	[2.731]	(0.371)	
CPCS Index[c]	27.700	27.004	0.696	832
	[2.876]	[3.217]	(0.391)	
Female	0.599	0.629	-0.030	843
	[0.491]	[0.484]	(0.030)	
Age	9.897	9.938	-0.042	839
	[1.108]	[1.193]	(0.304)	
Age Squared	99.166	100.186	-1.020	839
	[22.387]	[24.329]	(6.062)	

Notes: The sample consists of those who have at least the endline data. We replace the DT test 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 is zero are shown in parentheses and clustered at the school level. Superscripts ***, **, *, denote the statistical significance obtained by clustered wild bootstrap-t procedures at the 1, 5, and 10 percent levels, respectively.

[a]: DT stands for math Diagnostic Test. We use three outcomes of DT for measuring cognitive abilities: DT score, DT time, and DT score per minute (DT scores per min).

[b]: PTSII-C Score stands for math proficiency test scores.

[c]: The Proficiency Test of Self Learning is based on 27 survey questions, of which 10 are consistent with the Children's Perceived Competence Scale (CPCS Index) and 8 with the Rosenberg Self-Esteem Scale (RSES Index). For each noncognitive-type question, see Appendix C.

domization.[21] Note that the number of observation is smaller than the intended sample size discussed above because of attrition. In addition to attrition owing 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 DT. This is one of the main limitations of this study. Therefore, to maximize the sample size, we adopt the analysis of covariance (ANCOVA) specification in the main analysis (discussed in the subsequent section). We do so by replacing missing values with the mean

[21] As a robustness check, the sample with those who had both baseline and endline records of each variable are reported in Table D1 in Appendix D and also shows the baseline balance. Please note that the number of observation is smaller than the sample size in Table 1 because of attrition. In addition to attrition owing 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. Further, we drop observations if there is any missing in survey questions that comprise CPCS or RSES Indices.

of all non-missing baseline outcomes. In the estimation, we include a dummy variable that indicates that the baseline outcome is missing. Regarding noncognitive variables (i.e., RSES and CPCS Indices), some students could not answer any of the corresponding survey questions to construct the indices owing to time constraints. Therefore, we drop such cases from the analysis. This lead to slightly smaller sample size in comparison to that of the PTSII-C score. Consequently, the number of the observation in the final sample is 811 for the DT, 837 for the PTSII-C test, and 832 for the RSES and CPCS Indices. 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 including all and present the results in Tables J3 and J4 in Appendix J, which are qualitatively similar.

2.6 Sample Attrition

While some attrition emerges in our sample at the endline, the attrition rate is not significantly different between the treatment and control groups (Table D3 in Appendix D). 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 1 if s/he has the endline outcome and the value 0 if not.[22] Table D3 shows that the attrition did not systematically happen with respect to the treatment status and it would not cause the selection issues in our estimates.

3 Empirical Specification and Results

In the main analysis, we adopt the ANCOVA specification (McKenzie, 2012; Ma et al., 2020), in addition to the simple endline comparison. Let t denote the time period where $t = 0$ illustrates the baseline and $t = 1$ represents the endline. Let Y_{it} be a measure of cognitive or noncognitive abilities of student i at time t; d_i the treatment status (taking

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

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} error terms. If Y_{i0} is missing, we insert the mean value of the baseline into it. Then, the simple endline comparison is based on:

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

while the ANCOVA specification can be written as

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

Here, the average treatment effects on the treated can be captured by the estimated δ.[23] We use cluster robust standard errors at the school level. However, given the relatively smaller number of clusters, we use a wild cluster bootstrap procedure, following Cameron, Gelbach and Miller (2008).[24]

To investigate heterogeneous treatment effects, we estimate equation (2) for four different sub-samples: i) students with high initial cognitive ability and high initial noncognitive ability (high–high type); ii) students with high initial cognitive ability and low initial noncognitive ability (high–low type); iii) students with low initial cognitive ability and high initial noncognitive ability (low–high type); and iv) students with low initial cognitive ability and low initial noncognitive ability (low–low type). The cut-off points for high and low are the median values of the respective outcome measures at the baseline.[25]

[23] In the previous working paper (Sawada et al., 2020), we employ the canonical difference-in-differences (DID) model to estimate the impact of the Kumon intervention on the measures of cognitive and noncognitive abilities of student i at time t, Y_{it}: $Y_{it} = \alpha_0 + \alpha_1 T_t + \phi d_i + \delta^{did} T_t \cdot d_i + u_i + e_{it}$. Here, T_t is a time dummy taking 1 for endline 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:

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

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

[24] Unlike the standard cluster-robust standard errors, which are downward biased, this approach reduces over-rejection 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).

[25] 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

The parameters of interest are δ for different initial ability types.

3.1 Impacts 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 endline comparison based on Equation (1). Conversely, Panel B confirms these findings in Panel A with ANCOVA specification based on Equation (2). Note that all of the outcome variables are standardized so the magnitudes of the impacts are reported in their standard deviations.[26]

The first four columns of Table 2 shows 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 s.d. Effect size based on the ANCOVA specification is similar (0.465 s.d.) as illustrated in Panel B. Furthermore, as discussed above, time reduction in solving questions is the other important dimension in developing cognitive abilities. Therefore, we examine the treatment effects using the measures which 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 them to solve DT. The latter is the test score of PTSII-C, which has 228 questions, which is beyond the number 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.[27] Before examining 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

time are the outcome variables, while we use PTSII-C when PTSII-C and noncognitive abilities are the dependent variables.

[26] 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 using wild bootstrap.

[27] Furthermore, contrary to standard exams including DT, the students' scores of PTSII-C will not reach the full score. Therefore, this measurement partially avoids the typical censoring problem in estimating treatment effects. See Kawarazaki et al. (2022) for the detail of the discussion.

Table 2. Impact of Kumon on Students' Learning Outcomes

Dependent Variable	DT Score[a] (1)	DT Time[a] (2)	DT Score per min[a] (3)	PTSII-C Score[b] (4)	RSES Index[c] (5)	CPCS Index[c] (6)
Panel A: Endline Estimates						
Treatment	0.429***	-2.461***	2.283***	0.900***	0.086	0.176
	(0.128)	(0.426)	(0.406)	(0.208)	(0.150)	(0.145)
Constant	0.610***	-0.733	0.847*	0.859***	-0.052	-0.094**
	(0.106)	(0.228)	(0.143)	(0.126)	(0.085)	(0.084)
N	811	811	811	837	832	832
R-squared	0.080	0.267	0.204	0.147	0.002	0.007
p-value (individual hypothesis testing)	0.002	0.000	0.000	0.000	0.571	0.232
p-value (individual hypothesis testing, wild bootstrap)	0.002	0.002	0.000	0.000	0.561	0.218
Panel B: ANCOVA Estimates						
Treatment	0.465***	-2.209***	2.085***	0.999***	0.056	0.131
	(0.144)	(0.527)	(0.528)	(0.210)	(0.139)	(0.129)
Baseline Outcome	0.135***	0.046	0.295***	0.335***	0.107**	0.101**
	(0.049)	(0.123)	(0.095)	(0.083)	(0.049)	(0.040)
Constant	0.601***	-0.725***	0.837***	0.810***	0.026	-0.003
	(0.104)	(0.220)	(0.133)	(0.115)	(0.089)	(0.084)
N	811	811	811	837	832	832
R-squared	0.106	0.281	0.221	0.228	0.027	0.034
p-value (individual hypothesis testing)	0.003	0.000	0.000	0.000	0.687	0.314
p-value (individual hypothesis testing, wild bootstrap)	0.000	0.002	0.000	0.000	0.655	0.338

Notes: The sample is the same as that in Table 1. Panel A presents the result from the endline estimate based on Equation (1), while Panel B that from ANCOVA specification, which is based on Equation (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. Superscripts ***, **, and * denote the statistical significance obtained by clustered wild bootstrap-t procedures at the 1, 5, and 10 percent levels, respectively.

[a]: DT stands for math diagnostic test. We use three outcomes of DT for measuring cognitive abilities: DT score, DT time, and DT score per minute (DT scores per min).

[b]: PTSII-C score stands for math proficiency test scores.

[c]: The proficiency test of self-learning is based on 27 survey questions, of which 10 are consistent with the children's perceived competence scale (CPCS Index) and 8 with the Rosenberg self-esteem scale (RSES Index). For each of noncognitive-type question, see Appendix C.

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 impact is sizable: treatment effects measured by DT score per minute with Equation (2) is 2.085, as shown in Column (3) in Panel B. While this effect size may seem surprisingly high compared to the effect size of education interventions elsewhere, the effect size on DT score per minute is owing to a substantial reduction in test completion time measured as DT time (-2.209 s.d.), discussed above. Similarly, the treatment effects measured by PTSII-C with Equation (2) is 0.999, as shown in Column (4) in Panel B, partly reflecting the time-reduction effects. Note that effect size of the DT score (0.465 s.d.), that is, improvement in the raw test score, is consistent with previous findings in the literature wherein it is found to be effective in improving learning outcomes. Unlike previous studies that have used test scores to determine cognitive ability, we use test scores per minute (DT score per minute), as our intervention is designed to increase students' abilities to solve math problems in a time-efficient manner, an important ability in pursuing more

Table 3. Heterogeneous Impact of Kumon on Students' Learning Outcomes

	Initial RSES Index					Initial CPCS Index				
Dependent Variable	DT Score[a]	DT Time[a]	DT Score per min[a]	PTSII-C Score[b]	RSES Index[c]	DT Score[a]	DT Time[a]	DT Score per min[a]	PTSII-C Score[b]	CPCS Index[c]
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: High Initial Cognitive and High Initial Noncognitive Group[d]										
Treatment	0.326*	-3.477***	2.962***	1.123***	-0.031	0.303*	-3.236***	2.913***	1.042***	0.193
	(0.173)	(0.610)	(0.692)	(0.276)	(0.193)	(0.149)	(0.675)	(0.692)	(0.319)	(0.184)
Baseline Outcome	0.183	-0.114	-0.048	0.198*	-0.084	0.103	0.076	0.146	0.144	-0.072
	(0.145)	(0.398)	(0.383)	(0.111)	(0.114)	(0.195)	(0.521)	(0.624)	(0.148)	(0.127)
Constant	0.642***	-0.445	0.904***	0.847***	0.276*	0.738***	-0.731**	0.938*	0.914***	0.231
	(0.139)	(0.263)	(0.237)	(0.165)	(0.161)	(0.189)	(0.336)	(0.458)	(0.224)	(0.169)
N	159	159	159	189	189	149	149	149	188	188
R-squared	0.067	0.402	0.296	0.245	0.004	0.059	0.342	0.266	0.202	0.012
Panel B: High Initial Cognitive and Low Initial Noncognitive Group[d]										
Treatment	0.247	-3.459***	2.885***	1.135***	0.136	0.264	-3.614***	2.741***	1.251***	0.004
	(0.160)	(0.710)	(0.902)	(0.334)	(0.259)	(0.202)	(0.704)	(0.902)	(0.265)	(0.219)
Baseline Outcome	0.184	1.403	1.642***	0.222	0.349***	0.276	0.031	0.381	0.279***	0.223**
	(0.182)	(0.909)	(0.548)	(0.138)	(0.111)	(0.206)	(0.439)	(0.510)	(0.090)	(0.103)
Constant	0.705***	-0.839**	0.047	0.849***	0.280	0.598***	-0.404*	0.636*	0.781***	0.183
	(0.188)	(0.363)	(0.401)	(0.248)	(0.218)	(0.173)	(0.210)	(0.337)	(0.178)	(0.194)
N	124	124	124	180	180	134	134	134	181	181
R-squared	0.041	0.462	0.316	0.187	0.056	0.051	0.478	0.310	0.231	0.022
Panel C: Low Initial Cognitive and High Initial Noncognitive Group[d]										
Treatment	0.607**	-2.301***	2.457**	1.146***	0.108	0.506*	-2.071**	2.225**	0.980***	0.095
	(0.273)	(0.807)	(1.023)	(0.303)	(0.214)	(0.284)	(0.818)	(1.023)	(0.322)	(0.256)
Baseline Outcome	0.343***	-0.428**	0.647*	0.545**	0.181	0.210	-0.510***	0.592*	0.611**	0.102
	(0.120)	(0.182)	(0.371)	(0.211)	(0.190)	(0.145)	(0.159)	(0.340)	(0.284)	(0.242)
Constant	0.663***	-0.416	0.928***	0.784***	-0.171	0.658***	-0.441	0.999***	0.959***	-0.054
	(0.234)	(0.292)	(0.304)	(0.233)	(0.187)	(0.212)	(0.286)	(0.281)	(0.301)	(0.265)
N	118	118	118	150	150	109	109	109	150	150
R-squared	0.164	0.199	0.164	0.244	0.015	0.102	0.162	0.127	0.194	0.004
Panel D: Low Initial Cognitive and Low Initial Noncognitive Group[d]										
Treatment	0.540***	-1.410***	1.172***	0.980***	0.306*	0.620***	-1.599***	1.406***	1.110***	0.345*
	(0.170)	(0.498)	(0.333)	(0.219)	(0.160)	(0.158)	(0.476)	(0.333)	(0.234)	(0.184)
Baseline Outcome	0.113	1.203	0.019	0.870***	0.036	0.162**	1.550	-0.019	0.772***	-0.137
	(0.073)	(1.151)	(0.217)	(0.248)	(0.122)	(0.072)	(1.136)	(0.240)	(0.225)	(0.098)
Constant	0.705***	-0.908**	0.973***	1.195***	-0.228*	0.688***	-0.942**	0.852***	1.036***	-0.407***
	(0.127)	(0.370)	(0.232)	(0.187)	(0.116)	(0.130)	(0.419)	(0.207)	(0.160)	(0.109)
N	143	143	143	177	177	152	152	152	177	177
R-squared	0.174	0.132	0.109	0.290	0.021	0.217	0.171	0.150	0.315	0.036

Notes: The sample is the same as in Table 1. The estimation is based on the ANCOVA specification, which is illustrated in Equation (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. Superscripts ***, **, and * denote the statistical significance obtained by clustered wild bootstrap-t procedures at the 1, 5, and 10 percent levels, respectively.

[a]: DT stands for math diagnostic test. We use three outcomes of DT for measuring cognitive abilities: DT score, DT time, and DT score per minute (DT scores per min).

[b]: PTSII-C Score stands for math proficiency test scores.

[c]: The proficiency test of self-learning is based on 27 survey questions, of which 10 are consistent with the children's perceived competence scale (CPCS Index) and 8 with the Rosenberg self-esteem scale (RSES Index). For each of the noncognitive type-questions, see Appendix C.

[d]: The initial cognitive score stands for the baseline DT score for Columns (1)–(3), as well as Columns (6)–(8), and the baseline PTSII-C score for Columns (4),(5),(9), and (10). The initial noncognitive score stands for the baseline RSES Index for Columns (1)–(5). The baseline CPCS Index is used in Columns (6)–(10).

Table 4. Impact of Kumon on Students' Learning Outcomes: Estimates Controlling for Longer Sessions

Dependent Variable	DT Score (1)	DT Time (2)	DT Score per min[a] (3)	PTSII-C Score[b] (4)	RSES Index[c] (5)	CPCS Index[c] (6)
Treatment	0.419**	-2.580***	2.355***	1.002***	0.088	0.198
	(0.152)	(0.606)	(0.682)	(0.263)	(0.169)	(0.151)
Treatment × Longer session	0.132	1.069*	-0.778	-0.010	-0.089	-0.183
	(0.143)	(0.566)	(0.571)	(0.305)	(0.222)	(0.211)
Baseline Outcome	0.140***	0.031	0.264**	0.335***	0.104**	0.094**
	(0.047)	(0.129)	(0.099)	(0.082)	(0.050)	(0.040)
Constant	0.601***	-0.723***	0.837***	0.810***	0.024	-0.008
	(0.104)	(0.221)	(0.133)	(0.115)	(0.089)	(0.083)
N	811	811	811	837	832	832
R-squared	0.110	0.305	0.232	0.228	0.028	0.037
p-value (individual hypothesis testing)	0.010	0.000	0.002	0.001	0.606	0.199
p-value (individual hypothesis testing, wild bootstrap)	0.012	0.002	0.000	0.000	0.621	0.226

Notes: The sample is the same as that in Table 1. The estimation is based on the ANCOVA specification, which is illustrated in Equation (2) with the additional term for the interaction between the treatment and longer session dummies. 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. Superscripts ***, **, and * denote the statistical significance obtained by clustered wild bootstrap-t procedures at the 1, 5, and 10 percent levels, respectively.

[a]: DT stands for math diagnostic test. We use three outcomes of DT for measuring cognitive abilities: DT score, DT time, and DT score per minute (DT scores per min).

[b]: PTSII-C score stands for math proficiency test scores.

[c]: The proficiency test of self-learning is based on 27 survey questions, of which 10 are consistent with the children's perceived competence scale (CPCS Index) and 8 with the Rosenberg self-esteem scale (RSES Index). For each noncognitive-type question, see Appendix C.

complex materials in higher education. In contrast, regarding the noncognitive outcomes reported in the last two columns in Panel B of Table 2, the homogeneous treatment effect size estimates are insignificant.[28]

The heterogeneous treatment effects are reported in Panels A through 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 with high-initial cognitive abilities (high-high type and high-low type). However, they are smallest for students with low initial abilities in both measures (low-low type). Regarding noncognitive outcomes, however, we find suggestive evidence of the catch-up effect: students with 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 those with sufficiently improved noncognitive ability (i.e., higher impacts on cognitive ability compared to low-high type to low-low type).

As students in treatment schools have studied Kumon materials for an additional 30 minutes per day, one might argue that the impact estimates we present here may be attributed to longer session times in schools and not merely owing to the Kumon interven-

[28] 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 endline. As shown in Tables E1 and E2 in Appendix E together with the baseline balancing results on Tables J3 and J4, the impact 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 using the wild bootstrap. Lastly, in Table E2, the p-value (Romano–Wolf stepdown p-value) is reported based on 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).

tion. We investigate this possibility. Note that in the difference-in-differences (DID), the constant term shows improvement in control group, which means that baseline improvement with 60 minutes of study on mathematics. Therefore, when we subtract 1.5 times of the baseline improvement (60 plus 30 minutes) from the size of the treatment effect assuming the effect is constant to scale in terms of time, if there is some positive effect remaining, that would be the effects from Kumon program. According to Column (4) in Panel C of Table E2, which shows the DID results, the constant term, 0.839 s.d., is the improvement in DT score per minute for the control group by attending regular BRAC math classes only. If the impact of extending math learning hours is linear, 50 percent longer hours of learning math should be equivalent to 1.2585 s.d. (= 0.839 s.d.×1.5) worth of impacts measured in DT score per minute. If we subtract this longer study-hour effect size (1.2585 s.d.) from the treatment coefficient (2.073 s.d.), we have 0.8145 s.d. or 39.3 percent of the treatment effect. This remains to be a fairly large treatment effect.[29] Similarly, if we used the effect size of PTSII-C (1.212 s.d.) and subtract 50 percent longer study-hour effect size (0.679 s.d.×1.5), we have 0.1935 s.d. Although the number seems much smaller than that of DT score per minute, we still see sizable effects. In fact, the assumption on constant return to scale seems conservative. Figure F1 shows the average cumulative worksheets numbers along the cumulative Kumon session days. This shows how students have learned with the Kumon program. The Kumon learning curve is slightly concave, which indicates that the students' rate of improvement in math learning outcomes decreases as study hours lengthen. Hence, the back-of-the-envelope counterfactual calculation of longer study hours using the linear assumption might be conservative. Therefore, these numbers would be the lower bound of the treatment effects.[30] Furthermore, we exploit the fact that some treatment schools conducted Kumon sessions for at least 5 minutes longer. Using these time variations in the Kumon sessions, we examine the impact of the longer study time of Kumon (Table 4). Insignificant coef-

[29]This may suggest decreasing the return to scale of the standard lecture-style learning, which would also support the effectiveness of Kumon as a complementary program.

[30]However, it should be mentioned that when doing the balance test using the DID sample, the CPCS variable turned out to be marginally unbalanced (Table J4), which implies that we need to be little cautious in interpreting the findings.

ficients on the cross-term between the treatment and longer-session dummy suggest that overall outcomes are not systematically affected by longer school sessions. An additional 5 minutes did not change the treatment effects, which may be a result of the flattening learning curve (sharply decreasing marginal impact beyond 30 minutes). These results suggest that Kumon program itself contributes to the positive treatment effects.

3.2 Long-term Impact

To assess the long-term impact of the intervention, we use additional information from a national examination that certifies the completion of primary education (Primary School Certificate; PSC) after 8 and 20 months of the intervention, respectively, for the fourth and third grade students in our study. Specifically, we use information about the PSC examination take-up and dropouts and math scores obtained by students in our sample.[31]

First, we found that the PSC take-up rate is higher among students in treated schools (50.5 percent) than the rate among those in control schools (47.7 percent), albeit their statistically insignificant difference as shown in Table D4 in Appendix D.[32] Considering that only about a half of students took the PSC, 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 its completion time of the treatment school students is significantly lower than that of the control school students (Table D5). We show the distribution of the PSC Math letter grades of both treatment and control group students in Figure G1 in Appendix G. 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

[31] We collected students' PSC registration IDs from BPS branch offices and teachers at the schools. We then obtained their PSC results from government websites based on IDs. We also collected information from schools about dropouts from the PSC (non-takers).

[32] The primary reason for not taking the primary terminal examination was family relocation (79 percent). Conversely, other reasons included dropouts because of labor market participation (8.5 percent), school change (7.3 percent), early marriage (1.5 percent), sickness (0.75 percent), death (0.24 percent), and miscellaneous (2.7 percent). 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, 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. As previously mentioned, only letter grades are available for math grades.

Table 5. Long-Term Impact of Kumon on Students' Learning Outcomes

Panel A: OLS Results

Dependent Variable	First Difference of PSC Math Score and Baseline PTSII-C Score		
	(1)	(2)	(3)
Treatment	0.179	0.178	0.248
	(0.366)	(0.367)	(0.340)
Controlling for Gender		x	x
Controlling for Age, and Age Squared			x
N	445	445	445

Panel 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	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	0.226*	0.235*	0.244**	0.233**	0.040	0.051	0.320*	0.314*
	(0.123)	(0.120)	(0.121)	(0.118)	(0.157)	(0.165)	(0.165)	(0.164)
Constant			-0.270***	-0.240***				
			(0.095)	(0.090)				
Control in Selection Equation (Gender)						x		x
N	445	445	445	445				
N of Selected Obs.[a]					445	445	445	445
N of Total Obs.[b]					905	905	905	905

Notes: Panel A presents the result from the 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. Superscripts ***, **, and * denote the statistical significance obtained by clustered wild bootstrap-t procedures at the 1, 5, and 10 percent levels, respectively. In the analysis, we replace missing values for age with 0, but indicate a missing dummy in the estimation.

[a]: Number of observations whose record of the endline outcome is observable.

[b]: Number of total observations, including those without the record of the endline outcome.

Figure G2 in Appendix G, we show 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 to stay to take PSC than control students. This also may suggest that 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 existence of a sample selection problem among PSC exam participants that needs to be mitigated in our estimation. Accordingly, we employ quasi-experimental methods. To eliminate selection bias arising from the unobserved time-invariant heterogeneity, we employ four estimation models: difference-in-differences (DID), propensity score matching (PSM), inverse probability weighting (IPM), and Lee's Bounds methods.

To assess the long-term impact of the intervention, we employ the PSC math and PTSII-C test scores as the endline and baseline outcomes in a standardized form, respectively. First, apart from the sample selection problem, we undertake the standard difference-in-differences analysis using the difference between standardized PSC math score (endline) 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 5).[33] Second, to mitigate potential selection bias arising from the endogenous decision of taking the PSC exam, we also employ propensity score matching (PSM) and inverse probability weighting (IPW) methods where we match the sample based on pre-treatment student characteristics (i.e., student age, age squared, and a gender dummy).[34] As shown in Panel B of Table 5, results suggest that students from treatment schools received statistically significantly higher scores than those from control schools wherein point estimates of treatment effects range from 0.226

[33]To construct this outcome measure, we standardize the PSC and PTSII-C results and take their difference. However, this may not be an ideal outcome variable because, by construction, the measures are different. This is a potential caveat, but standardization partially mitigates the problem of comparability between PSC and PTSII-C.

[34]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.

s.d. to 0.244 s.d.[35] Comparison of the OLS estimation results in Panel A with the results from PSM method and IPW regression in Panel B of Table 5 suggests that endogenous selection in taking PSC might have generated downward bias in estimating treatment effects.[36] Third, we also estimate the Lee's bounds (Lee, 2009) consider nonrandom sample selection in taking the PSC exam with the monotonicity assumption that is no heterogeneous effect of treatment on selection. As shown in Panel B of Table 5, upper bound estimates are statistically significant. Point estimates of PSM and IPW (Panel B) are within these Lee's bounds (0.040 s.d. to 0.320 s.d.). Overall, we find a modest but positive long-term impact of the intervention on cognitive ability measured by math test score. Moreover, we show the heterogenous treatment effects by the baseline PTS-II score (Figure H1 in Appendix H):[37] Treatment effects seem higher for students whose baseline PTSII-C scores are in the 40–60 and 80–100 percentiles.[38] Although estimation is imprecise, most students benefited from Kumon intervention.[39]

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 PSC are different, in addition to the several age cohorts in our analysis. Therefore, we conduct the heterogeneous analyses in terms of age and initial ability. Figure H2 and Table H1 in Appendix H show the results. Estimation results suggest that treatment effects are higher and positive when the intervention is done when students are young. Conversely, the effects gradually fade out and become negative when it is done when they are old. This pattern is consistent with literature on educational intervention saying that a childhood intervention should be conducted as early as possible (Heckman, 2006; Hendren and Sprung-Keyser, 2020) and the effects might be deteriorating when it is too late (Chetty, Hendren and Katz, 2016).

[35] See Footnote 18 for the PSC grading scale.

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

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

[38] The average baseline PTSII-C score is not statistically different between the treatment and control school students who took the PSC exam. Standardized mean of PTSII-C scores among PSC takers at the treatment schools (-0.105) and control schools (0.334) differ by 0.439 (p-value: 0.110). However, the standardized mean of DT score of PSC takers from the treatment schools (-0.021) and control school (0.266) significantly differ by 0.287 (p-value: 0.088).

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

Furthermore, we also examine heterogeneity in terms of initial grade. As in Figure H1, the better the initial ability, the larger the long-term impacts.

3.3 Teacher Assessment Ability

In addition to student outcomes, we examine the impact 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 student' abilities as 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 6, 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 impacts 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 worksheets and speed of 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 them with better information about their students.[40]

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

Table 6. Association between Teacher's Assessment and Student Performance

Dependent Variable	Absolute Difference between Teacher's Perception and Student's Score	
	DT Score[a]	PTSII-C Score[b]
	(1)	(2)
Treatment × Endline	-0.919***	-0.350**
	(0.265)	(0.132)
Treatment	-0.045	-0.219
	(0.294)	(0.142)
Endline	-0.248	0.148*
	(0.192)	(0.077)
Constant[c]	2.346***	1.535***
	(0.241)	(0.110)
N	990	1416
R-squared	0.101	0.047

Notes: The dependent variable is the absolute difference between the teacher's subjective evaluation and student's objective performance. Asymptotic standard errors are shown in parentheses and clustered at the school level. Superscripts ***, **, and *, denote the statistical significance obtained by clustered wild bootstrap-t procedures at the 1, 5, and 10 percent levels, respectively.
[a]: DT Score per min stands for math Diagnostic Test scores per minute.
[b]: PTSII-C Score stands for math proficiency test scores.
[c]: The significance level of the coefficients is based on the standard p-value.

4 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 impact estimate on math PSC scores (Table 5) and estimated wage returns to numeracy skills from Nordman, Sarr and Sharma (2015) that use the matched employer-employee data. Benefit per student is calculated as a product of the impact of Kumon on math ability (s.d.), wage returns on numeracy skills (s.d.), and average annual earnings.[41] We assume that the benefit will last from 1 to 44 years, considering working age as 16 to 59 and an annual discount rate of five percent, following Duflo (2001). We did not use the

[41]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 5. Wage returns to numeracy skills, 0.037, are taken from Table 3, column 8 of Nordman, Sarr and Sharma (2015). Average annual earnings are calculated based on average hourly wage in Table 2 of Nordman, Sarr and Sharma (2015) (50.91), multiplied by 40 hours per week and 52 weeks. We then calculate the life cycle profile of earnings based on estimates of the returns to tenure and tenure-squared in Nordman, Sarr and Sharma (2015)'s regression (0.037 and -0.00067).

Dead-weight loss factor as this program did not involve tax spending or revenue.

As the minimum cost, we consider worksheet printing costs based on 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 percent higher worksheet printing costs if some students completed a higher level, regardless of use. According to 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 one when benefits last for more than 14 (more than 16) years, as shown in Figure I1 (Figure I2) in Appendix I. However, it should be noted that the wage returns to numeracy skills are estimated based on full-time formal sector jobs, which is a growing sector but not necessarily a representative type of employment in Bangladesh. IRR is calculated so the present values of benefit and cost equalize over a specified time horizon, varying from 1 year to 44 years. IRR becomes positive when workers continue working with benefits for more than 10 (11) years with the minimum (maximum) cost (Figures I1 and I2).

5 Discussion

In this study, we have investigated 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 have conducted a field experiment to test the effectiveness of the Kumon mathematics learning program in improving primary school students' cognitive and noncognitive abilities in Bangladesh. 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 8 months. This was offered among BRAC primary school 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 impact using time-adjusted test scores, whereby impact 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 to education interventions elsewhere. Our relatively larger impact estimates may be attributed to the extra math learning over a period of 8 months as opposed to personalized learning. This is because we observe substantially higher impacts in test score per minute compared to raw test score, implying that students become more time-efficient in solving math problems. Extra studying time effects could account for as much as 61 percent of our impact estimates, and the rest can be attributed to the personalized and broader learning effects. Moreover, the intervention is particularly designed to improve math problem-solving skills through building endurance and perseverance. Hence, we find 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. Particularly, our results show some long-term impact 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 BPS teacher's role during Kumon session was limited to monitoring and mechanically determining the level of worksheets based on the predefined procedure, we have found positive impacts on BPS teachers capacity to assess student performance. This finding implies that BPS teachers may have benefited from Kumon intervention by gaining more objective information about students' skills. However, we have no evidence suggesting that intervention affected their regular teaching practice. Future research should focus on teachers' perception and teaching practice.

This paper-and-pencil-based self-learning program is well-suited for the setting constrained by inadequacy of ICT infrastructure and therefore easily scalable in developing

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

We note here some potential limitations of our analysis. Firstly, some observations were dropped owing to non-compliance and attrition resulting in smaller sample size than 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 might be useful to enhance the external validity of our results. Secondly, our long-term impact analysis is based on public examinations on national curriculum administered after 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 owing to family relocation issues. However, no significant difference was found in PSC take-up rate between treatment and control school students. We address potential selection bias using quasi-experimental analysis. Thirdly, in our benefit-cost analysis, we rely on PSC test scores after 8 and 20 months showing the long-term benefit of intervention. These impacts are already reduced compared to our main results. Therefore, the long-term benefit estimates shown here could potentially diminish over time. Finally, considering its focus, the current paper does not detail the mechanisms behind the impact of the Kumon method. In a companion paper, for example, we investigate the peer effects on classroom learning among treatment students (Kawarazaki et al., 2020). Uncovering these mechanisms would be a key task for future research.

References

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

Ahmad, Alia, and Iftekharul Haque. 2011. *Economic and Social Analysis of Primary Education in Bangladesh: A Study of BRAC Interventions and Mainstream Schools.* Vol. 48, BRAC Centre.

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(4): 432–452.

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

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(1): 75–106.

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." Cambridge, MA: National Bureau of Economic Research NBER Working Paper 22746.

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(3): 1235–1264.

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(3): 414–427.

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(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(4): 407–424.

Duflo, Esther. 2001. "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment." *American Economic Review*, 91(4): 795–813.

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

Evans, David K., and Anna Popova. 2015. "What Really Works to Improve Learning in Developing Countries?: An Analysis of Divergent Findings in Systematic Reviews." Washington, DC: World Bank World Bank Policy Research Working Paper 7203.

Ganimian, Alejandro J., and Richard J. Murnane. 2016. "Improving Education in Developing Countries: Lessons From Rigorous Impact Evaluations." *Review of Educational Research*, 86(3): 719–755.

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

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

Heckman, James J. 2006. "Skill Formation and the Economics of Investing in Disadvantaged Children." *Science*, 312(5782): 1900–1902.

Heckman, James J. 2007. "The Economics, Technology, and Neuroscience of Human Capability Formation." *Proceedings of the National Academy of Sciences*, 104(33): 13250–13255.

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

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(3): 411–482.

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(1-2): 114–128.

Hendren, Nathaniel, and Ben Sprung-Keyser. 2020. "A Unified Welfare Analysis of Government Policies." *The Quarterly Journal of Economics*, 135(3): 1209–1318.

Kawarazaki, Hikaru, Minhaj Mahmud, Yasuyuki Sawada, and Mai Seki. 2020. "Haste Makes No Waste: Positive Peer Effects of Speed Competition on Classroom Learning." Center for International Research on the Japanese Economy CIRJE Working Paper CIRJE-F-1151.

Kawarazaki, Hikaru, Minhaj Mahmud, Yasuyuki Sawada, Mai Seki, and Kazuma Takakura. 2022. "Still Biased? A Remaining Classical Selection Problem of RCTs in Education." Center for International Research on the Japanese Economy CIRJE Working Paper CIRJE-F-1194.

Kremer, Michael, Conner Brannen, and Rachel Glennerster. 2013. "The Challenge of Education and Learning in the Developing World." *Science*, 340(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(7)(e65775): 1–13.

Lee, David S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *Review of Economic Studies*, 76(3): 1071–1102.

Ma, Yue, Robert W. Fairlie, Prashant Loyalka, and Scott Rozelle. 2020. "Isolating the "Tech" from EdTech: Experimental Evidence on Computer Assisted Learning in China." National Bureau of Economic Research Working Paper 26953.

McEwan, Patrick J. 2015. "Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments." *Review of Educational Research*, 85(3): 353–394.

McKenzie, David. 2012. "Beyond Baseline and Follow-up: The Case for More T in Experiments." *Journal of Development Economics*, 99(2): 210–221.

Muralidharan, Karthik, Abhijeet Singh, and Alejandro J. Ganimian. 2019. "Disrupting education? Experimental evidence on technology-aided instruction in India." *American Economic Review*, 109(4): 1426–1460.

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

Nath, Samir Ranjan. 2015. *Whither Grade V Examination? An assessment of primary educational completion examination in Bangladesh.*

Nordman, Christophe J., Leopold R. Sarr, and Smriti Sharma. 2015. "Cognitive, Non-Cognitive Skills and Gender Wage Gaps: Evidence from Linked Employer-Employee Data in Bangladesh." DIAL (Développement, Institutions et Mondialisation) Working Papers DT/2015/19.

Romano, Joseph P., and Michael Wolf. 2005. "Stepwise Multiple Testing as Formalized Data Snooping." *Econometrica*, 73(4): 1237–1282.

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

Sabarwal, Shwetlena, Malek Abu-Jawdeh, and Radhika Kapoor. 2022. "Teacher Beliefs: Why They Matter and What They Are." *The World Bank Research Observer*, 37(1): 73–106.

Sakurai, Shigeo, and Yutaka Matsui, ed. 1992. *Shinri Sokutei Shakudo Shu IV (Psychological measurement scale IV): Jido-you Konnpitensu Shakudo (Competence scale for children) "Jikokachi (Self-Worth)".* 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." CIRJE Discussion Papers CIRJE-F-1127, University of Tokyo.

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

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

Watkins, Kevin. 2000. *The Oxfam Education Report.* Oxford: Oxfam International.

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

Online Appendix

A Kumon Method Worksheet Examples

In the Kumon method, the self-learning process is enforced by examples and hints (the first few questions with gray lines). Furthermore, students only need to learn new math concepts and calculation steps in very small increments on each worksheet, helping them learn autonomously. For example, the first worksheet (3A1a) allows students to learn the order of numbers (up to 100). Once students have mastered these worksheets without errors within a targeted time frame, they begin to learn the concept of addition (note: completion within a targeted time is a proxy for permitting students to advance to the next worksheet). The second worksheet (3A71a) introduces students to the concept of "adding 1," using just an arrow. This concept follows from the number order list that students have already mastered before reaching this level. Finally, in the third worksheet (3A74a), students learn the concept of adding 1 using the summation sign (i.e., "+ 1").

The final worksheet (D81a) shows division by two-digit numbers. Even with more complicated arithmetic, the examples and hints and preceding worksheets allow students to self-learn calculation skills and some of the math concepts behind them. Please note that these worksheets comprise the English versions thereof. For the BRAC primary school trail, all materials were translated into Bengali, the local language BRAC primary school students regularly use in class.

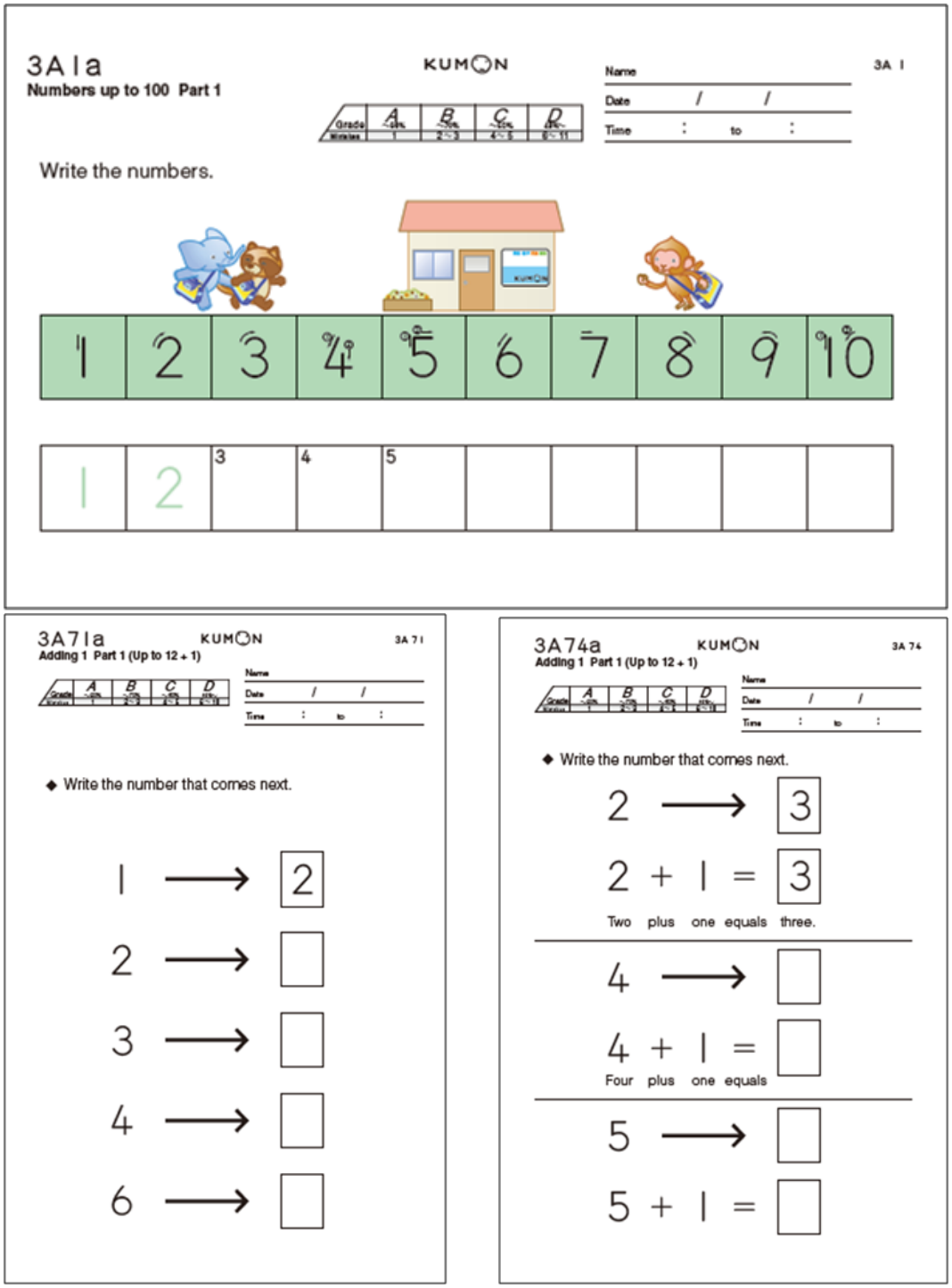
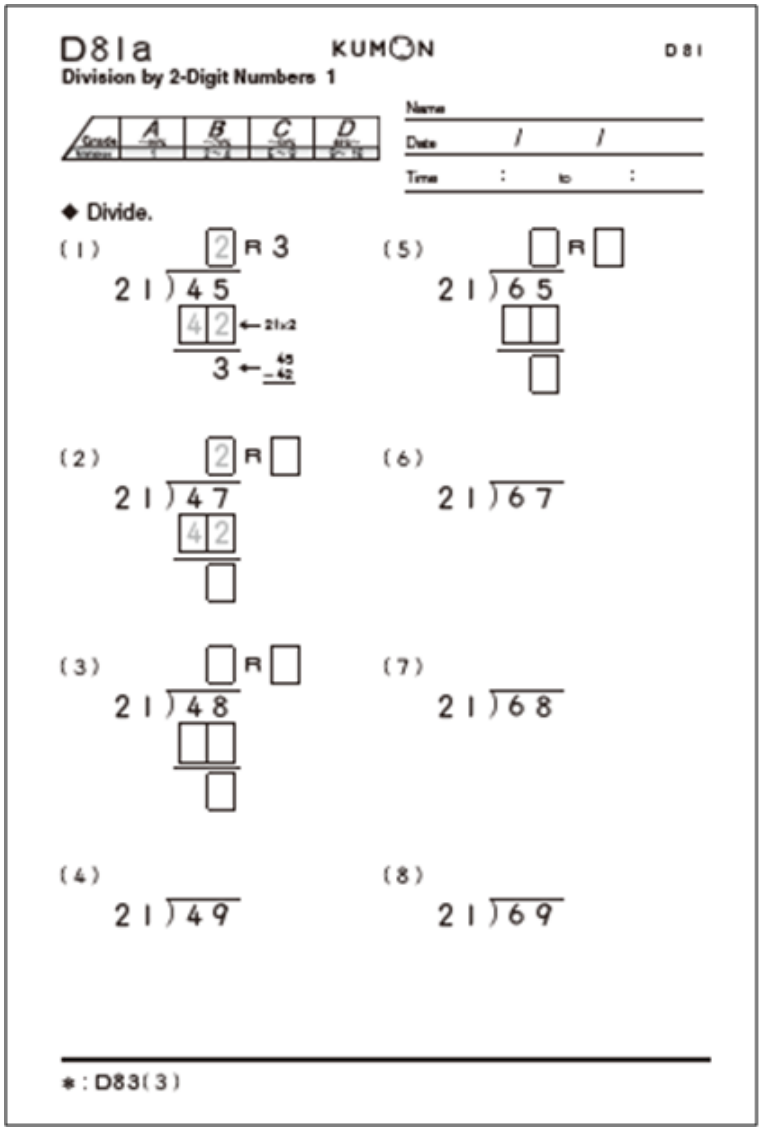


Figure A1. Examples of Problem-Solving Math Worksheets



D81a KUMON D81

Division by 2-Digit Numbers 1

Name

Date / /

Time : to :

◆ Divide.

(1) 2 R 3
21)45
42 ← 21×2
3 ← 45 − 42

(2) 2 R □
21)47
42
□

(3) □ R □
21)48
□□
□

(4) 21)49

(5) □ R □
21)65
□□
□

(6) 21)67

(7) 21)68

(8) 21)69

★ : D83(3)

Figure A2. Examples of Problem-Solving Math Worksheets (Cont.)

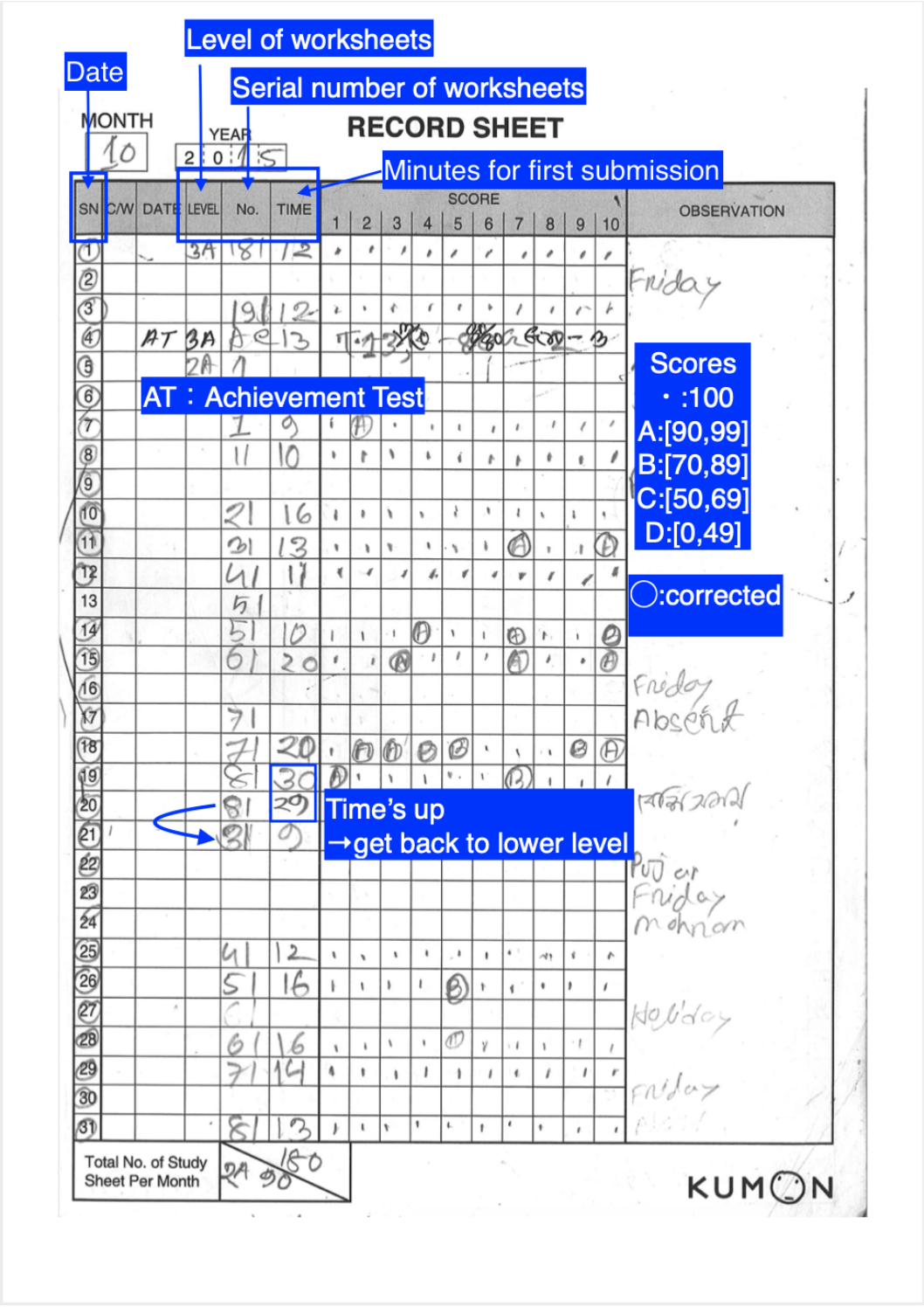
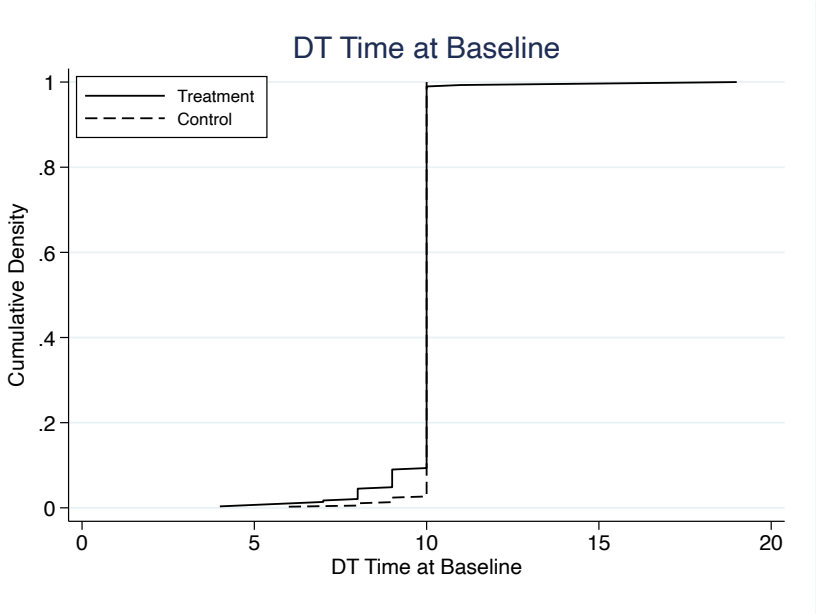


Figure A3. Example of a Record Sheet in a Record Book

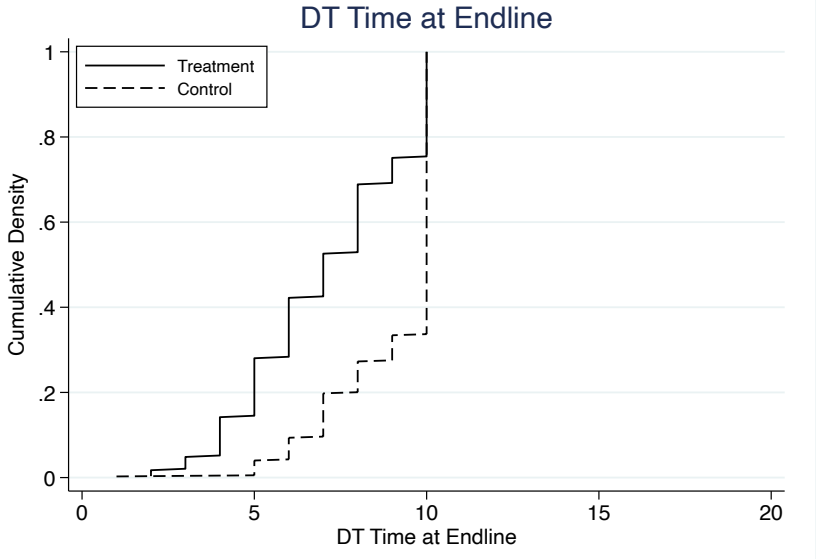
B DT Time Management

Figure B1. Cumulative Density Functions of the DT Time at Baseline



Notes: The sample for this figure includes only those whose baseline and endline on DT time were recorded for comparison. We also exclude observations with the wrong DT.

Figure B2. Cumulative Density Functions of the DT Time at Endline



Notes: The sample for this figure includes only those whose baseline and endline on DT time were recorded for comparison. We also exclude the observations with wrong DT.

C Noncognitive Ability Survey Questions

Table C1. PTS II Survey Questions for Measuring Noncognitive Abilities

Number	Question in English	RSES	CPCS
1	I did well on this test.		
2	I can do most things better than others.	x	x
3	There are many things about myself I can be proud of.	x	x
4	I feel that I cannot do anything well no matter what I do.	x	x
5	I believe I can be someone great.		x
6	I don't think I am a helpful person.	x	x
7	I can confidently express my opinion.		x
8	I don't think I have that many good qualities.	x	x
9	I am always worried that I might fail.	x	x
10	I am confident in myself.	x	x
11	I am satisfied with myself.	x	x
12	Even if I fail, I think I can get better and better at things if I keep trying.		
13	I like to do calculations.		
14	I can calculate in my head when I go shopping.		
15	I think speed is important when solving problems.		
16	While studying, I believe everything will go well if I correctly follow the instructions.		
17	I am more motivated when people praise me.		
18	I always volunteer in class.		
19	I enjoy studying.		
20	School is fun.		
21	I do things better when I have a goal.		
22	There are many things I want to learn more about.		
23	a. I have a role model around me. b. There is someone who I want to be like.		
24	I always have someone who I can go to for advice when I am having trouble with my studies.		
25	a. There is someone who I do not want to lose against. b. There is someone who I am always competing with.		
26	I always try to do something when things don't go as expected.		
27	It doesn't matter whether I fail in the beginning because I believe that things will eventually work out.		

Note: Among 27 questions, we use 8 of the 10 full questions of the Rosenberg Self-Esteem Scale (RSES) (Rosenberg, 1965), and 10 full questions of the Children's Perceived Competence Scale (CPCS) (Sakurai and Matsui, 1992; Harter, 1979). The rest are more specific to the original Kumon method of learning with four Bangladesh-specific questions (questions 24–27). The Japanese version of the original Kumon survey questions was based on Sakurai and Matsui (1992).

Table C2. Level of Kumon Worksheets

	Level	Sheet Number	Contents
Highest	F	2001–2200	Addition, subtraction, multiplication, and division of fractions
	E	1801–2000	Addition of fractions
	D	1601–1800	Column division
	C	1401–1600	Column multiplication
	B	1201–1400	Column addition
	A	1001–1200	Subtraction based on mental arithmetic
	2A	801–1000	Addition based on mental arithmetic
	3A	601–800	Addition based on number tables
	4A	401–600	Writing numbers and understand the order of numbers
	5A	201–400	Counting numbers up to 50
Lowest	6A	1–200	Counting numbers from one to ten

Note: In each level, we have 200 worksheets. We convert the difficulty level of worksheet into numerical values, using sheet numbers from 1-200 (lowest level) to 2001-2200 (highest level).

D Sample Attrition

This appendix documents the robustness checks of the attrition status and related balancing test results. First, Table D1 reports balancing test results based on the minimum sample, which includes those who have a complete record of both the baseline and endline outcomes and excludes those who took the wrong DT or those who miss even one component of noncognitive indices (RSES or CPCS). Consequently, sample size is smaller than the sample used in the main ANCOVA analysis. However, even with this attrition, the baseline is balanced.

Second, in Table D2, we check the balance between two groups, including those who took the DT of a wrong level in the treatment group, in addition to those without baseline outcomes with values were replaced with the baseline mean. As the level was not adjusted, the questions might be too easy for them. Therefore, DT time will be reasonably shorter. This possibility seems to reflect the negative significant sign in DT time comparison. However, once we control the dummy for the wrong DT, the difference disappears. This suggests that, although randomisation might not be perfect, the effects from this incompleteness would be minimal. In the main analysis, we treat the baseline records of these observations as missing to adopt the simplest ANCOVA specification without additional control variables. Results of the main findings do not change quantitatively if we do not replace them with the missing value and do control for the dummy variable indicating wrong DT assignment.

Third, Table D3 shows whether attrition status of the main sample correlates with any outcome variables. As we adopt the ANCOVA strategy and the missing baseline is replaced with mean of the baseline values, which will not be attributed to the sample attrition, the sample size consists of the number of all the observations. Dependent variables are dummy variables that take the value 1 if the corresponding endline records are missing and 0 if otherwise. As established, attrition status between the treatment and control groups do not have any systematic differences.

Fourth, as our analysis is extended to examining long-term effects, we conduct at-

trition status on the long-term outcome, PSC examination results. As described in the main text, we document significant dropouts before the PSC examination, and baseline imbalance would be the potential problem. First, Table D4 shows the result of baseline balancing of the sample between PSC takers and non-takers (external margin). Next, we examine the potential difference in baseline outcomes within the PSC between treatment and control students (internal margin). The Table D5 shows baseline imbalance in DT outcomes by the treatment status, suggesting the need for adopting a quasi-experimental approach, such as propensity score matching (PSM) and Inverse Probability Weighting (IPW), while controlling for potential selection that would arise from time-invariant characteristics using the DID specification. Once we use the PSM approach, as shown in Table D6 for the matched sample, we do not see the baseline imbalance. This suggests the importance of correction and validity of the main analysis with a quasi-experimental approach.

Table D1. Baseline Balance Test Results with Strictly Balanced Panel Data

Dependent Variable	Treatment	Control	Difference	N
DT Score	47.419	47.291	0.127	663
	[15.608]	[16.555]	(2.944)	
DT Time	9.879	9.960	-0.081	663
	[0.918]	[0.295]	(0.072)	
DT Score per min[a]	4.894	4.757	0.137	663
	[1.943]	[1.693]	(0.322)	
PTSII-C Score[b]	34.665	39.040	-4.375	787
	[10.603]	[15.508]	(3.666)	
RSES Index[c]	20.915	21.022	-0.107	371
	[3.093]	[3.195]	(0.546)	
CPCS Index[c]	27.841	26.994	0.846	360
	[3.458]	[3.858]	(0.547)	
Female	0.599	0.629	-0.030	843
	[0.491]	[0.484]	(0.030)	
Age	9.897	9.938	-0.042	839
	[1.108]	[1.193]	(0.304)	
Age Squared	99.166	100.186	-1.020	839
	[22.387]	[24.329]	(6.062)	

Notes: The sample comprises those who have information on both baseline and endline data. We treat the DT test results of those who took the wrong DT as missing. Standard deviations are presented in brackets. Asymptotic standard errors based on testing the hypotheses that differences between the treatment and control is zero are shown in parentheses and clustered at the school level. Superscripts ***, **, *, denote the statistical significance obtained by clustered wild bootstrap-t procedures at the 1, 5, and 10 percent levels, respectively.

[a]: DT stands for math Diagnostic Test. We use three outcomes of DT for measuring cognitive abilities: DT score, DT time, and DT score per minute (DT scores per min).

[b]: PTSII-C Score stands for math proficiency test scores.

[c]: The Proficiency Test of Self Learning is based on 27 survey questions, of which 10 are consistent with the Children's Perceived Competence Scale (CPCS Index) and 8 with the Rosenberg Self-Esteem Scale (RSES Index). For each noncognitive type question, see Appendix C.

Table D2. Baseline Balance Test Results (ANCOVA Sample) with Wrong DT as Missing

Panel A: Balance Test

Dependent Variable	Treatment	Control	Coefficient	N
DT Score	46.116	47.267	-1.151	811
	[17.416]	[16.402]	(2.942)	
DT Time	9.453	9.956	-0.503**	811
	[1.396]	[0.294]	(0.204)	
DT Score per min[a]	5.128	4.759	0.369	811
	[2.562]	[1.678]	(0.372)	

Panel B: Regression Result with the Dummy for the Wrong DT

Dependent Variable	Coefficient	N
DT Score	0.137	811
	(2.890)	
DT Time	-0.078	811
	(0.071)	
DT Score per min[a]	0.135	811
	(0.317)	

Notes: Standard deviations are shown in brackets. Asymptotic standard errors are shown in parentheses and are clustered at the school level.
[a]: DT Score per min stands for math Diagnostic Test scores per minute.
[b]: PTSII-C Score stands for math proficiency test scores.
[c]: The Proficiency Test of Self Learning is based on 27 survey questions, of which 10 are consistent with the Children's Perceived Competence Scale (CPCS Index) and 8 with the Rosenberg Self-Esteem Scale (RSES Index). For each noncognitive-type question, see Appendix C.

Table D3. Attrition Status

Dependent Variable	Attrition Status across Outcome Measures		
	DT[a]	PTSII-C Score[b]	RSES/CPCS Index[c]
	(1)	(2)	(3)
Treatment	-0.020	0.070	0.063
	(0.072)	(0.046)	(0.048)
Constant	0.203***	0.130***	0.138***
	(0.062)	(0.030)	(0.032)
Num of Obs.	1004	1004	1004
R-squared	0.001	0.009	0.007

Notes: Asymptotic standard errors are shown in parentheses and are clustered at the school level. The superscripts, ***, **, *, denote the statistical significance obtained by clustered wild bootstrap-t procedures at the 1, 5, and 10 percent levels, respectively.
[a]: DT stands for math Diagnostic Test. Attrition status among DT Score, Time, and DT Score per Minute are identical.
[b]: PTSII-C Score stands for math proficiency test scores.
[c]: The Proficiency Test of Self Learning is based on 27 survey questions, of which 10 are consistent with the Children's Perceived Competence Scale (CPCS Index) and 8 with the Rosenberg Self-Esteem Scale (RSES Index). For each noncognitive-type question, see Appendix C.

Table D4. PSC Take-up

Dependent Variable	Treatment	Control	Difference	N
PSC Take-up	0.505	0.477	0.028	905
	[0.501]	[0.500]	(0.075)	

Notes: The sample consists of those who have at least the baseline data of PTSII. Standard deviations are presented in brackets. The column for difference shows the regression coefficient of the treatment dummy where we regress the dummy variable for the PSC take-up on the treatment dummy. 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. Superscripts ***, **, and * denote the statistical significance obtained by clustered wild bootstrap-t procedures at the 1, 5, and 10 percent levels, respectively.

Table D5. Baseline Balance for PSC Takers

Dependent Variable	Treatment	Control	Coefficient	N
DT Score	44.928	50.534	-5.606*	442
	[18.070]	[15.222]	(2.922)	
DT Time	9.373	9.951	-0.579*	442
	[1.383]	[0.353]	(0.271)	
DT Score per min[a]	5.053	5.095	-0.042	442
	[2.647]	[1.591]	(0.425)	
PTSII-C Score[b]	35.412	41.507	-6.095	445
	[10.483]	[14.769]	(4.007)	
RSES Index[c]	21.031	21.081	-0.148	445
	[2.660]	[2.670]	(0.402)	
CPCS Index[c]	27.830	27.098	0.648	445
	[3.091]	[3.234]	(0.471)	
Female	0.655	0.652	0.003	445
	[0.476]	[0.477]	(0.053)	
Age	10.123	9.957	0.166	443
	[1.125]	[1.133]	(0.302)	
Age Squared	103.733	100.411	3.322	443
	[23.065]	[23.198]	(6.124)	

Notes: The sample consists of those who have at least PSC record. We treat the DT test results of those who took wrong DT as missing. This is different from Table 1, because we adopt the difference in differences specification for PSC analysis. Standard deviations are shown in brackets. The column for Coefficient shows the regression coefficient of treatment dummy where we regress each dependent variable on treatment dummy and missing dummy for cognitive and noncognitive outcomes. Asymptotic standard errors based on testing the hypotheses that the differences between the treatment and control is zero are shown in parentheses and are clustered at the school level. The superscripts, ***, **, *, denote the statistical significance obtained by clustered wild bootstrap-t procedures at the 1, 5, and 10 percent levels, respectively.
[a]: DT Score per min stands for math Diagnostic Test scores per minute.
[b]: PTSII-C Score stands for math proficiency test scores.
[c]: The Proficiency Test of Self Learning is based on 27 survey questions, of which 10 are consistent with the Children's Perceived Competence Scale (CPCS Index) and 8 with the Rosenberg Self-Esteem Scale (RSES Index). For each of the noncognitive type questions, see Appendix C.

Table D6. Baseline Balance for PSC Takers (Matched Sample in PSM)

Dependent Variable	Treatment	Control	Difference	N
DT Score	46.090	51.547	-5.457	323
	[15.995]	[14.018]	(3.063)	
DT Time	9.962	9.947	0.015	323
	[0.542]	[0.367]	(0.057)	
DT Score per min[a]	4.654	5.200	-0.545	323
	[1.673]	[1.480]	(0.327)	
PTSII-C Score[b]	35.594	41.675	-6.080	420
	[10.700]	[14.859]	(4.104)	
RSES Index[c]	21.046	21.088	-0.041	402
	[2.695]	[2.753]	(0.442)	
CPCS Index[c]	27.853	27.077	0.776	402
	[3.143]	[3.334]	(0.517)	
Female	0.657	0.652	0.005	443
	[0.476]	[0.477]	(0.053)	
Age	10.123	9.957	0.166	443
	[1.125]	[1.133]	(0.302)	
Age Squared	103.733	100.411	3.322	443
	[23.065]	[23.198]	(6.124)	

Notes: The sample consists of those who remained after matching in PSM regression, whereby we match the sample based on pre-treatmentstudent characteristics (i.e., student age, age squared, and gender). Standard deviations are shown in brackets. Asymptotic standard errors are shown in parentheses and clustered at the school level.

[a]: DT Score per min stands for math Diagnostic Test scores per minute.

[b]: PTSII-C Score stands for math proficiency test scores.

[c]: The Proficiency Test of Self Learning is based on 27 survey questions, of which 10 are consistent with the Children's Perceived Competence Scale (CPCS Index) and 8 with the Rosenberg Self-Esteem Scale (RSES Index). For each noncognitive-type question, see Appendix C.

E Robustness Analysis

This section provides the main result with different specifications and samples: (i) result from the same sample as Table 2, except we omit from the sample those who took the wrong DT (Table E1); and (ii) result from the sample for difference-in-difference specification as conducted in our previous working paper (Sawada et al., 2020) (Table E2).[42] Tables with odd numbers show the balancing test results for the sample, of which regression results are shown in tables with successive even numbers.

As in Tables E1 and E2, the effects of the Kumon intervention are robust to different specifications.

[42] We show the corresponding balancing checks on Tables J3 and J4 in Appendix J.

Table E1. Impact of Kumon on Students' Learning Outcomes (ANCOVA Sample Without Those with Wrong DT)

Dependent Variable	DT Score	DT Time	DT Score per min[a]	PTSII-C Score[b]	RSES Index[c]	CPCS Index[c]
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Endline Estimates						
Treatment	0.480***	-2.189***	2.079***	0.900***	0.086	0.176
	(0.138)	(0.549)	(0.545)	(0.208)	(0.150)	(0.145)
Constant	0.610***	-0.733	0.847*	0.859***	-0.052	-0.094**
	(0.106)	(0.228)	(0.143)	(0.126)	(0.085)	(0.084)
Num of Obs.	673	673	673	837	832	832
R-squared	0.092	0.209	0.179	0.147	0.002	0.007
p-value (individual hypothesis testing)	0.002	0.000	0.001	0.000	0.571	0.232
p-value (individual hypothesis testing, wild bootstrap)	0.000	0.002	0.000	0.000	0.561	0.218
Panel B: ANCOVA Estimates						
Treatment	0.484***	-2.188***	2.073***	0.999***	0.056	0.131
	(0.140)	(0.546)	(0.544)	(0.210)	(0.139)	(0.129)
Baseline Outcome	0.136***	0.047	0.295***	0.335***	0.107**	0.101**
	(0.049)	(0.123)	(0.095)	(0.083)	(0.049)	(0.040)
Constant	0.592	-0.734	0.843	0.810	0.026	-0.003
	(0.101)	(0.225)	(0.133)	(0.115)	(0.089)	(0.084)
Num of Obs.	673	673	673	837	832	832
R-squared	0.122	0.209	0.191	0.228	0.027	0.034
p-value (individual hypothesis testing)	0.002	0.000	0.001	0.000	0.687	0.314
p-value (individual hypothesis testing, wild bootstrap)	0.000	0.002	0.000	0.000	0.655	0.338

Notes: Asymptotic standard errors are presented in parentheses and are clustered at the school level. Superscripts ***, **, and * denote the statistical significance obtained by clustered wild bootstrap-t procedures at the 1, 5, and 10 percent levels, respectively.
[a]: DT score per min stands for math diagnostic test scores per minute.
[b]: PTSII-C score stands for math proficiency test scores.
[c]: The proficiency test of self-learning is based on 27 survey questions, of which 10 are consistent with the children's perceived competence scale (CPCS Index) and 8 with the Rosenberg self-esteem scale (RSES Index). For each of the noncognitive-type question, see Appendix C.

Table E2. Impact of Kumon on Students' Learning Outcomes (DID Sample Excluding Those with Wrong DT)

Dependent Variable	DT Score (1)	DT Time (2)	DT Score per min[a] (3)	PTSII-C Score[b] (4)	RSES Index[c] (5)	CPCS Index[c] (6)
Panel A: Endline Estimates						
Treatment	0.490***	-2.203***	2.103***	0.925***	0.120	0.179
	(0.137)	(0.552)	(0.548)	(0.212)	(0.160)	(0.149)
Constant	0.600***	-0.722***	0.831***	0.859***	-0.010	-0.031
	(0.104)	(0.226)	(0.133)	(0.124)	(0.099)	(0.089)
Num of Obs.	663	663	663	787	696	696
R-squared	0.095	0.211	0.182	0.152	0.003	0.007
p-value (individual hypothesis testing)	0.001	0.000	0.001	0.000	0.458	0.241
p-value (individual hypothesis testing, wild bootstrap)	0.000	0.002	0.000	0.000	0.476	0.266
p-value (Romano-Wolf stepdown p-value)	0.000	0.000	0.000	0.000	0.248	0.020
Panel B: ANCOVA Estimates						
Treatment	0.492***	-2.199***	2.094***	1.022***	0.110	0.151
	(0.140)	(0.551)	(0.549)	(0.209)	(0.155)	(0.145)
Baseline Outcome	0.136***	0.046	0.295***	0.337***	0.105**	0.100**
	(0.049)	(0.123)	(0.095)	(0.084)	(0.048)	(0.040)
Constant	0.589***	-0.729***	0.834***	0.798***	-0.002	-0.013
	(0.101)	(0.225)	(0.131)	(0.112)	(0.094)	(0.089)
Num of Obs.	663	663	663	787	696	696
R-squared	0.122	0.211	0.193	0.235	0.013	0.017
p-value (individual hypothesis testing)	0.002	0.000	0.001	0.000	0.484	0.303
p-value (individual hypothesis testing, wild bootstrap)	0.000	0.002	0.000	0.000	0.507	0.352
Panel C: First Difference Estimates						
Treatment	0.501**	-2.122***	2.073***	1.212***	0.026	-0.095
	(0.226)	(0.544)	(0.570)	(0.292)	(0.185)	(0.173)
Constant	0.521***	-0.881***	0.839***	0.679***	0.067	0.148
	(0.142)	(0.227)	(0.158)	(0.212)	(0.108)	(0.112)
Num of Obs.	663	663	663	787	696	696
R-squared	0.048	0.182	0.168	0.193	0.000	0.001
p-value (individual hypothesis testing)	0.035	0.001	0.001	0.000	0.891	0.588
p-value (individual hypothesis testing, wild bootstrap)	0.028	0.002	0.000	0.000	0.889	0.593
p-value (Romano-Wolf stepdown p-value)	0.000	0.000	0.000	0.000	0.851	0.485

Notes: Asymptotic standard errors are presented in parentheses and are clustered at the school level. Superscripts ***, **, and * denote the statistical significance obtained by clustered wild bootstrap-t procedures at the 1, 5, and 10 percent levels, respectively.
[a]: DT score per min stands for math diagnostic test scores per minute.
[b]: PTSII-C score stands for math proficiency test scores.
[c]: The proficiency test of self-learning is based on 27 survey questions, of which 10 are consistent with the children's perceived competence scale (CPCS Index) and 8 with the Rosenberg self-esteem scale (RSES Index). For each of noncognitive-type question, see Appendix C.

F Heterogeneity in Learning Speed

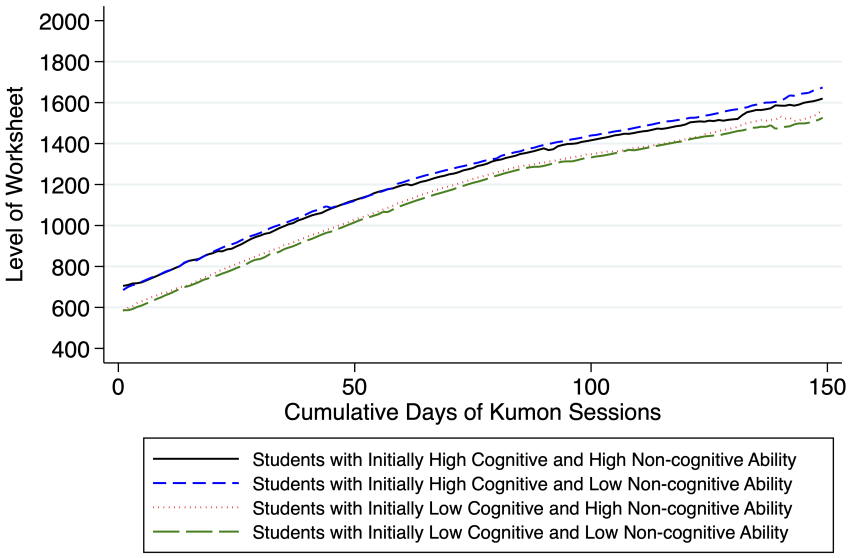


Figure F1. Heterogeneity in Learning Curve with Kumon Worksheets

Note: Levels of worksheet are converted to integers by combining alphabetical levels (6A to O) and number of worksheets. See Table F1 for details.

Table F1. Level of Kumon Worksheets

	Level	Sheet Number	Contents
Highest	F	2001–2200	Addition, subtraction, multiplication, and division of fractions
	E	1801–2000	Addition of fractions
	D	1601–1800	Column division
	C	1401–1600	Column multiplication
	B	1201–1400	Column addition
	A	1001–1200	Subtraction based on mental arithmetic
	2A	801–1000	Addition based on mental arithmetic
	3A	601–800	Addition based on number tables
	4A	401–600	Writing numbers and understand the order of numbers
	5A	201–400	Counting numbers up to 50
Lowest	6A	1–200	Counting numbers from one to ten

Note: In each level, we have 200 worksheets. We convert the difficulty level of worksheet into numerical values, using sheet numbers from 1-200 (lowest level) to 2001-2200 (highest level).

G Graphical Evidence of Math GPA from PSC

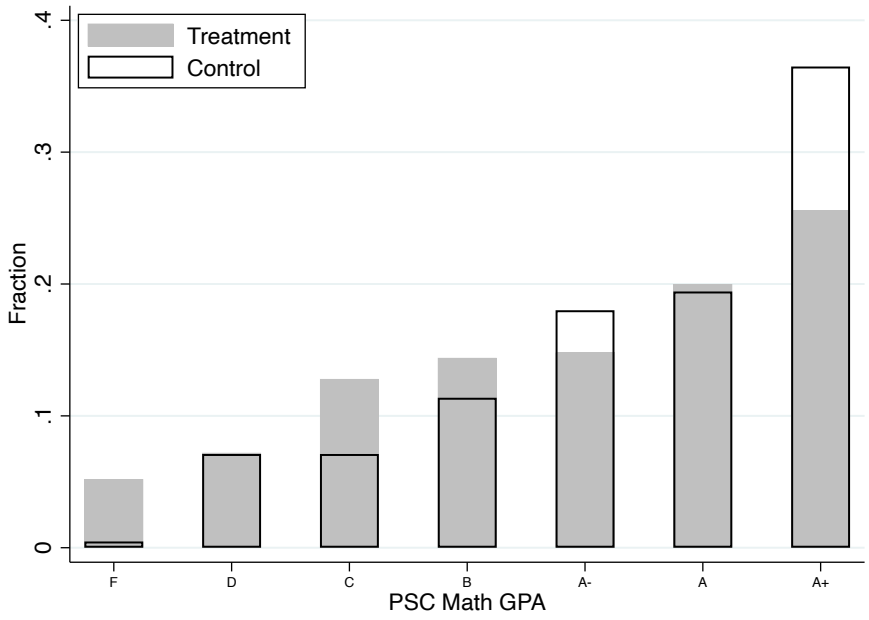


Figure G1. Histogram of Math GPA from PSC

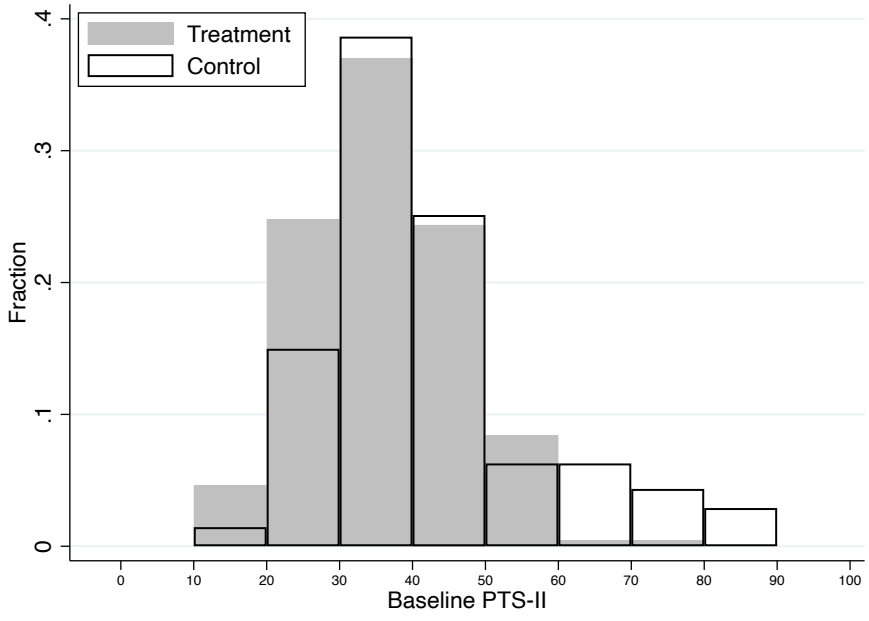


Figure G2. Histogram of Baseline PTS-II for PSC Takers

H Heterogeneous Effects on PSC Results

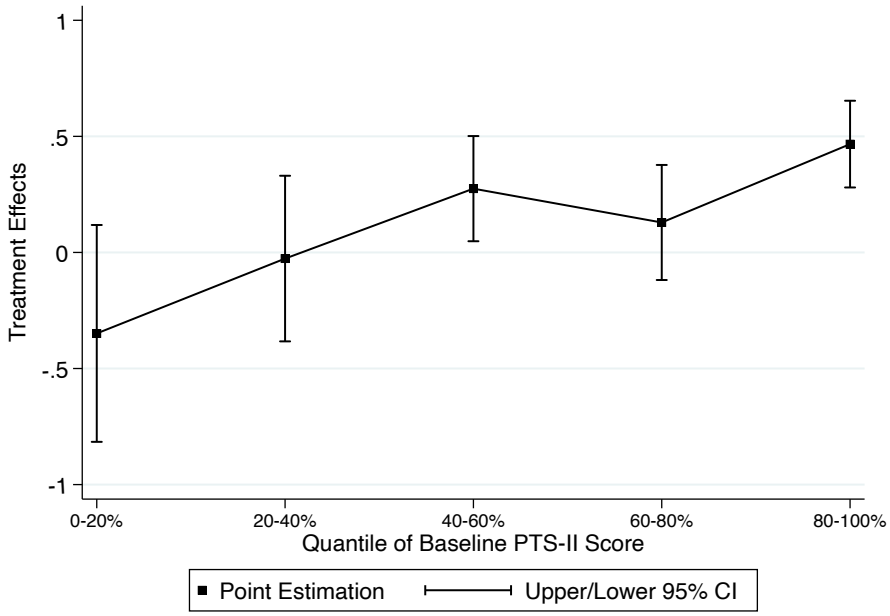


Figure H1. Heterogeneous Effects on PSC Results

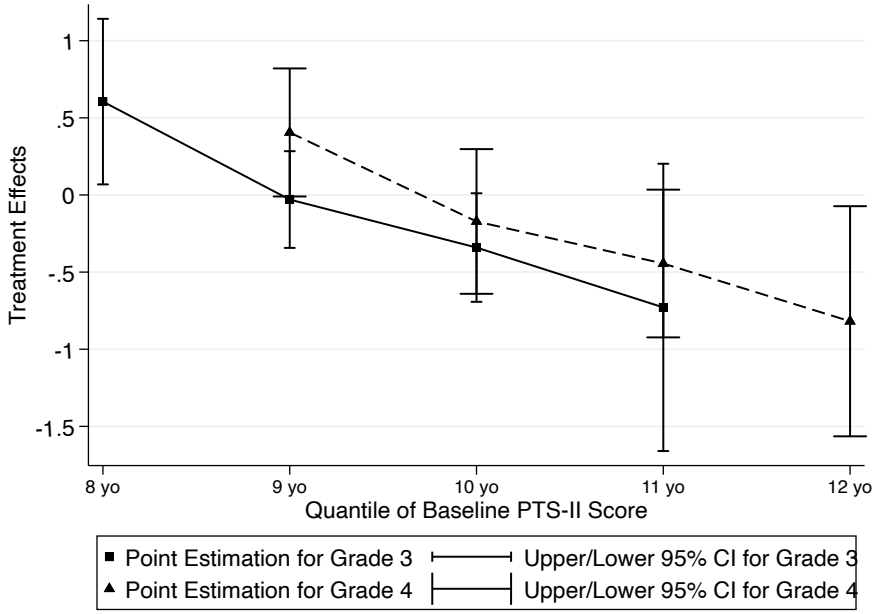


Figure H2. Heterogeneous Effects on PSC Results in Terms of Age

Note: We omit observations of ages 13 and 14 because of small sample size.

Table H1. Long-Term Impact of Kumon on Students' Learning Outcomes: Grade Heterogeneity

Panel A: Grade 3

	PSM regression		IPW regression		Lee bound			
	ATT estimates	ATE estimates	ATT estimates	ATE estimates	Lower Bound Estimate		Upper Bound Estimate	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	0.274*	0.319**	0.347**	0.355**	0.124	0.129	0.474**	0.455**
	(0.152)	(0.152)	(0.159)	(0.153)	(0.194)	(0.194)	(0.203)	(0.192)
Constant			-0.292**	-0.271**				
			(0.129)	(0.122)				
Control in Selection Equation (Gender)						x		x
Num of Obs.	226	226	226	226				
Num of Selected Obs.[a]					226	226	226	226
Num of Total Obs.[b]					512	512	512	512

Panel B: Grade 4

	PSM regression		IPW regression		Lee bound			
	ATT estimates	ATE estimates	ATT estimates	ATE estimates	Lower Bound Estimate		Upper Bound Estimate	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	0.191	0.224	0.166	0.176	-0.066	-0.102	0.203	0.191
	(0.196)	(0.185)	(0.191)	(0.183)	(0.239)	(0.229)	(0.240)	(0.241)
Constant			-0.287**	-0.229*				
			(0.142)	(0.129)				
Control in Selection Equation (Gender)						x		x
Num of Obs.	219	219	219	219				
Num of Selected Obs.[a]					219	219	219	219
Num of Total Obs.[b]					393	393	393	393

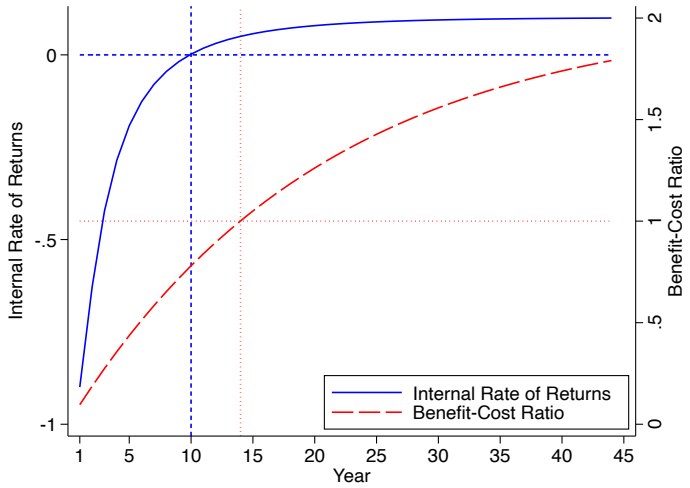
Notes: Panel A presents the result of Grade 3, while Panel B that of Grade 4. The specification is the same as in Panel B of Table 5. Asymptotic standard errors based on testing the hypotheses that the differences between treatment and control are zero are presented in parentheses. Superscripts ***, **, and * denote the statistical significance obtained by clustered wild bootstrap-t procedures at the 1, 5, and 10 percent levels, respectively. In this analysis, if data in the variables of age are missing (and therefore age squared), we substitute them with 0 and make the dummy variable indicating the substitution. In all regressions that control for age and age squared, we also control for this dummy variable. Therefore, contamination does not occur due to missing data.

[a]: Number of observations whose record of the endline outcome is observable.

[b]: Number of total observations, including those without the record of the endline outcome.

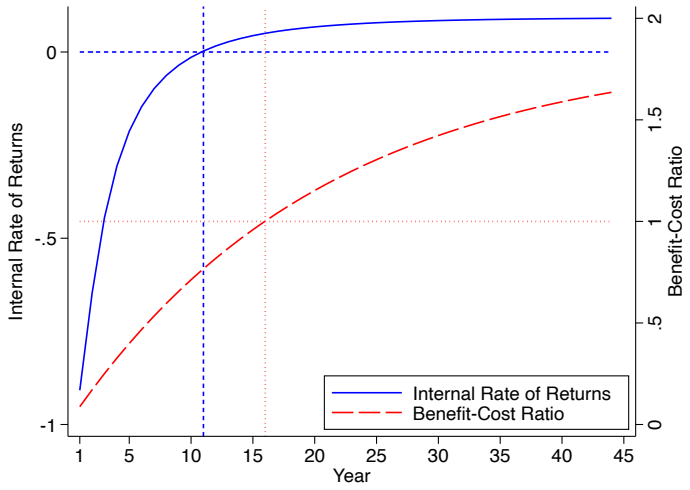
I Graphical Evidence of Benefit-Cost Analysis

Figure I1. Benefit-Cost (B-C) Ratio and Internal Rate of Return (IRR) with Minimum Cost



Notes: The blue solid line indicates internal rate of return (IRR), and the red long-dashed line indicates the benefit-cost ratio (BC). Blue dashed line indicates IRR = 0 and year = 10, while the red dotted line shows BC = 1 and year = 14.

Figure I2. Benefit-Cost (B-C) Ratio and Internal Rate of Return (IRR) with Maximum Cost



Notes: The blue solid line indicates internal rate of return (IRR). The red long-dashed line indicates benefit-cost ratio(BC). The blue dashed line indicates IRR = 0 and year = 11. Conversely, the red dotted line shows BC = 1 and year = 16.

J Additional Robustness Checks

This appendix presents tables on further robustness checks. Table J1 shows whether attrition status correlates with any outcome variables. This analysis is analogous to Table D3; however, the difference is in the samples used. In Table D3, we begin with the entire sample as we adopt the ANCOVA strategy, and the missing baseline will not be attributed to the sample attrition. In Table J1, however, we begin with the sample with baseline outcomes, which corresponds to the DID specification examined in Appendix E. As seen in Column (2) of Panel A in Table J1, we see the difference in the take-up rate in PTSII-C. However, as presented in Panel B, we do not see any difference in the outcome level, which suggests the validity of the analysis based on the DID approach.

Table J2 shows the balancing test result similar to Table D2. Here, those with a wrong DT are included as they are. However, the difference from Table D2 is that Table J2 does not include those without baseline records, which basically correspond to the sample for DID specification. However, this implication is the same as that of Table D2: we see the difference in take-up rate in PTSII-C, but this disappears once we control for the dummy indicating wrong DT distribution.

Tables J3 and J4, we report balancing test result on DID specification, corresponding to Appendix E we see almost no systematic differences between the treatment and control groups, which suggest the randomization was successful, except for the baseline CPCS, which warns us to more carefully interpret the effects.

Table J1. Attrition Status

Panel A: Sample Attrition

Dependent Variable	Attrition Status across Outcome Measures		
	DT[a]	PTSII-C Score[b]	RSES/CPCS Index[c]
	(1)	(2)	(3)
Treatment	0.060	0.096*	0.087
	(0.066)	(0.050)	(0.054)
Constant	0.169***	0.081**	0.095**
	(0.052)	(0.032)	(0.037)
Num of Obs.	825	905	812
R-squared	0.006	0.020	0.015

Panel B: Attrition Only Sample

Dependent Variable	Baseline PTSII-C Score
Treatment	-1.549
	(3.641)
Constant	35.657***
	(3.314)
Num of Obs.	118
R-squared	0.005

Notes: Asymptotic standard errors are shown in parentheses and clustered at the school level. Superscripts ***, **, and *, denote the statistical significance obtained by clustered wild bootstrap-t procedures at the 1, 5, and 10 percent levels, respectively.

[a]: DT stands for math Diagnostic Test. Attrition status among DT Score, Time, and DT Score per Minute are identical.

[b]: PTSII-C Score stands for math proficiency test scores.

[c]: The Proficiency Test of Self Learning is based on 27 survey questions, of which 10 are consistent with the Children's Perceived Competence Scale (CPCS Index) and 8 with the Rosenberg Self-Esteem Scale (RSES Index). For each noncognitive-type question, see Appendix C.

Table J2. Baseline Balance Test Results (ANCOVA Sample)

Panel A: Balance Test

Dependent Variable	Treatment	Control	Coefficient	N
DT Score	46.118	47.291	-1.174	799
	[17.519]	[16.555]	(2.989)	
DT Time	9.449	9.960	-0.510**	799
	[1.403]	[0.295]	(0.206)	
DT Score per min[a]	5.131	4.757	0.374	799
	[2.577]	[1.693]	(0.378)	

Panel B: Regression Result with the Dummy for the Wrong DT

Dependent Variable	Coefficient	N
DT Score	0.127	799
	(2.937)	
DT Time	-0.081	799
	(0.072)	
DT Score per min[a]	0.137	799
	(0.322)	

Notes: Standard deviations are shown in brackets. Asymptotic standard errors are shown in parentheses and are clustered at the school level.
[a]: DT Score per min stands for math Diagnostic Test scores per minute.
[b]: PTSII-C Score stands for math proficiency test scores.
[c]: The Proficiency Test of Self Learning is based on 27 survey questions, of which 10 are consistent with the Children's Perceived Competence Scale (CPCS Index) and 8 with the Rosenberg Self-Esteem Scale (RSES Index). For each noncognitive-type question, see Appendix C.

Table J3. Baseline Balance (ANCOVA sample excluding those with wrong DT)

Dependent Variable	Treatment	Control	Difference	N
DT Score	47.404	47.267	0.137	673
	[15.528]	[16.402]	(2.896)	
DT Time	9.877	9.956	-0.078	673
	[0.913]	[0.294]	(0.071)	
DT Score per min[a]	4.894	4.759	0.135	673
	[1.933]	[1.678]	(0.318)	
PTSII-C Score[b]	34.815	38.940	-4.124	837
	[10.191]	[15.195]	(3.489)	
RSES Index[c]	20.997	20.878	0.120	832
	[2.506]	[2.731]	(0.371)	
CPCS Index[c]	27.700	27.004	0.696	832
	[2.876]	[3.217]	(0.391)	
Female	0.599	0.629	-0.030	843
	[0.491]	[0.484]	(0.030)	
Age	9.897	9.938	-0.042	839
	[1.108]	[1.193]	(0.304)	
Age Squared	99.166	100.186	-1.020	839
	[22.387]	[24.329]	(6.062)	

Notes: Standard deviations are shown in brackets. Asymptotic standard errors are shown in parentheses and clustered at the school level.
[a]: DT Score per min stands for math Diagnostic Test scores per minute.
[b]: PTSII-C Score stands for math proficiency test scores.
[c]: The Proficiency Test of Self Learning is based on 27 survey questions, of which 10 are consistent with the Children's Perceived Competence Scale (CPCS Index) and 8 with the Rosenberg Self-Esteem Scale (RSES Index). For each noncognitive-type question, see Appendix C.

Table J4. Baseline Balance (DID sample excluding those with wrong DT)

Dependent Variable	Treatment	Control	Difference	N
DT Score[a]	47.419	47.291	0.127	663
	[15.608]	[16.555]	(2.944)	
DT Time[a]	9.879	9.960	-0.081	663
	[0.918]	[0.295]	(0.072)	
DT Score per min[a]	4.894	4.757	0.137	663
	[1.943]	[1.693]	(0.322)	
PTSII-C Score[b]	34.665	39.040	-4.375	787
	[10.603]	[15.508]	(3.666)	
RSES Index[c]	21.000	20.854	0.146	696
	[2.696]	[3.038]	(0.443)	
CPCS Index[c]	27.741	26.901	0.840*	696
	[3.092]	[3.571]	(0.468)	
Female	0.602	0.629	-0.026	829
	[0.490]	[0.484]	(0.029)	
Age	9.894	9.938	-0.044	829
	[1.118]	[1.193]	(0.307)	
Age Squared	99.141	100.186	-1.044	829
	[22.607]	[24.329]	(6.114)	

Notes: Standard deviations are shown in brackets. Asymptotic standard errors are shown in parentheses and are clustered at the school level.

[a]: DT stands for math Diagnostic Test. DT Score per min stands for math Diagnostic Test scores per minute.

[b]: PTSII-C Score stands for math proficiency test scores.

[c]: The Proficiency Test of Self Learning is based on 27 survey questions, of which 10 are consistent with the Children's Perceived Competence Scale (CPCS Index) and 8 with the Rosenberg Self-Esteem Scale (RSES Index). For each noncognitive-type question, see Appendix B.