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

By

Yasuyuki Sawada (The University of Tokyo and Asian Development Bank)
Minhaj Mahmud (Bangladesh Institute of Development Studies and JICA Research Institute)
Mai Seki (Ritsumeikan University)
An Le (NextGeM Inc.)
Hikaru Kawarazaki (The University of Tokyo)

October 2019

CREPE DISCUSSION PAPER NO. 60

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

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

Yasuyuki Sawada, Minhaj Mahmud, Mai Seki, An Le, and Hikaru Kawarazaki*

October, 2019

This paper investigates the effectiveness of a globally popular method of self-learning at the right level in improving the cognitive and non-cognitive abilities of disadvantaged pupils in a developing country, Bangladesh. Using a randomized control trial design, we find substantial improvement in cognitive ability measured by mathematics test scores and catch-up effects on non-cognitive ability measured by a pupil self-esteem measure. These findings are consistent with a longer-term impact found in take-up rates and scores on a national-level primary school completion exam. Moreover, the teachers' ability to assess student performance substantially improves. Based on our estimates, program benefit exceeds cost in a plausible way. Above findings suggest that self-learning at right level can effectively address the learning crisis by improving the quality of primary education in developing countries.
JEL: I20, O12
Keywords: education, self-learning, cognitive and non-cognitive outcomes, developing countries, randomized control trial

* This is 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." Sawada: Corresponding author. University of Tokyo, 7 Chome-3-1 Hongo, Bunkyo, Tokyo 113-8654, Japan, and Asian Development Bank (phone: +81-3-5841-5572, fax: +81-3-5841-5521, e-mail: sawada@e.u-tokyo.ac.jp); Mahmud: Bangladesh Institute of Development Studies, E-17 Agargaon, Sher-e- Bangla Nagar, Dhaka-1207, Bangladesh, and JICA Research Institute (e-mail: minhaj@bids.org.bd); Seki: Ritsumeikan University, 1-1-1 Nojihigarhi, Kusatsu, Shiga, 525-8577 Japan (e-mail: maisekijp@gmail.com); Le: NextGeM Inc., 6-3-5 Minatojima Minamimachi, Chuo-ku, Kobe-shi, Hyogo, 650-0047 Japan (e-mail: an.le@nextgem.jp); Kawarazaki: University of Tokyo, 7 Chome-3-1 Hongo, Bunkyo, Tokyo 113-8654, Japan (e-mail: hikaru-kawara@g.ecc.u-tokyo.ac.jp). The opinions expressed in this article are the authors' own and do not reflect the views of affiliated organizations. The research protocol was approved by the University of Tokyo IRB registered as No.15-90.

I. Introduction

Global successes have been recorded in terms of school enrollment as envisaged in the millennium development goals (MDGs). According to UNESCO (2015), there are 83 million fewer out-of-school children and adolescents as of 2012 than there were in 1999.[1] However, more than 60 percent of primary school children in low- and middle-income countries fail to achieve a minimum proficiency in mathematics and reading (World Bank, 2018; UNESCO, 2013). This crisis in learning is a serious concern among policy makers. Given that education is an important link to all the sustainable development goals (SDGs), improving the quality of education is a *sine qua non* for achieving them (United Nations, 2018). In this context, programs that match teaching to students' ability level and learning are gaining increased attention due to their high effectiveness in improving learning outcomes (Banerjee et al., 2007, 2016; Duflo, Dupas and Kremer, 2011; Muralidharan, Singh and Ganimian, 2019).[2] Nevertheless, the quality of teachers could prove a binding constraint for scaling up such interventions in a sustainable manner. Individualized self-learning programs can effectively address this constraint and thereby improve learning quality.

In Bangladesh, we test the effectiveness of a globally popular individualized program of self-learning to address the learning crisis. To improve the quantity of education, Bangladesh has been successful in increasing school enrollment and narrowing the gender gap. In this process, not only publicly provided education but also non-formal education has played a critical role. On the non-formal side, BRAC, the largest NGO in Bangladesh, has played a leading role in a collaboration with the government. In particular, BRAC primary schools (BPSs) have provided disadvantaged students with a four-year accelerated program that covers the five-year public primary school curriculum.[3] Given the success of BPS in terms of enrollment and reducing primary school dropouts, the government of Bangladesh has scaled up a modified version of BPS under the Reaching Out of School (ROSC) project, providing a low-cost platform to target children from difficult-to-reach communities and who are out of school (Asadullah, 2016). Despite these efforts, the lack of quality education and resulting inadequate student

[1] To achieve universal primary education in developing countries, a variety of policy interventions have been proposed and experimented with on both the supply and demand sides. These range from the expansion and improvement of school infrastructure to providing various incentives such as de-worming students, information sharing, free school lunches, free school uniforms, and conditional cash transfers (Kremer, 2003; Miguel and Kremer, 2004; Jensen, 2010; Duflo and Kremer, 2005; Banerjee and Duflo, 2006; Duflo, Glennerster and Kremer, 2007; Glewwe, 2002).

[2] In improving learning outcomes, demand-side approaches appear to be less promising than supply-side interventions such as increasing the numbers of teachers and schools. See Asim et al. (2017) for a meta-analysis of impact evaluation studies focusing on improving learning outcomes in South Asian countries. Other reviews focusing 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)

[3] BPS has been known as one of the largest and most successful non-formal education programs targeted to disadvantaged populations in Bangladesh. BPSs have introduced a seasonally adjusted school calendar, which has been a key to their success (Watkins, 2000; Chowdhury, Jenkins and Nandita, 2014). More details about BPS are discussed in section 2.

learning remain a serious concern in Bangladesh, as in other developing countries.[4] In this context, we adopt and evaluate the impact of the Kumon method of learning (hereafter Kumon) in improving both the cognitive and non-cognitive abilities of BPS students in Bangladesh, given its unique setting in providing non-formal education and internal efficiency compared to formal schools (Ahmad and Haque, 2011).[5] Kumon is a non-formal education program designed to ensure that each student always studies at a level that is "just right" for him/her.[6] This philosophy is similar to the "teaching at the right level (TaRL)" approach of Banerjee et al. (2016), although it also emphasizes the self-learning aspects of education. In Kumon, each student begins at an individually suitable starting point and learns new concepts in small steps where learning is enforced through easily understandable hints and examples. BPSs have 30 students per class with quite diverse backgrounds and a large variance in ability in the subjects taught, particularly mathematics (Nath, 2012). This creates a potential mismatch between teaching level and individual student ability. However, BPSs cannot effectively offer TaRL as they follow the same instructional approach as government schools such as lecture style education and "teaching to the test," potentially affecting learning outcomes. The Kumon program at least partially solves such a mismatch and improves learning outcomes by providing self-learning materials for each student in mathematics. Moreover, since the Kumon method of learning is based on a paper-and-pencil method, unlike the successful e-learning or computer-assisted instructions elsewhere (Banerjee et al., 2007; Barrow, Markman and Rouse, 2009; Muralidharan, Singh and Ganimian, 2019), the method is not constrained by limited or unstable electricity supplies we often experience in low-income communities of developing countries (United Nations, 2018).[7]

To preview our findings, Kumon has been found to substantially improve students' cognitive ability. Given that our intervention was designed to increase students' math problem-solving skills in a time-efficient manner, we use both test scores per minute and time-unadjusted test scores from two different mathematics

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

[5]While a number of existing studies have established the link between measured cognitive ability (e.g., IQ) and educational outcomes such as schooling attainment and wages, recent studies have begun to shed new light on the role of non-cognitive abilities such as personality traits, motivations, and preferences (Heckman, 2006, 2007). In fact, recent studies have begun to demonstrate that in explaining education, success in the labor market, or other outcomes, the predictive power of non-cognitive abilities is comparable to or exceeds that of cognitive skills (Heckman, 2006; Heckman, Humphries and Kautz, 2014). Notwithstanding this, Kumon has been regarded as a successful non-formal education program in strengthening both cognitive and non-cognitive outcomes, so it is worth evaluating its impacts in a disadvantaged environment where BPS has been operating.

[6]As of March 2017, there are 4.35 million subject enrollments in 50 countries and regions, according to the Kumon Institute of Education Co., Ltd.

[7]According to United Nations (2018), 13% of the global population still do not have access to modern electricity and three billion people still rely on traditional power sources for daily lives, such as wood, coal, charcoal, or animal waste for cooking and heating.

tests as measures of cognitive ability. The magnitude of the impact measured by test score per minute is a 2.177 standard deviation, whereby the impact comes through both test score gains and reductions in problem-solving speed.[8] In the case of the time-unadjusted test scores, the magnitude of the impact ranges from a 0.505 to a 1.198 standard deviation. In terms of non-cognitive abilities, we find catch-up effects among the pupils with initially low non-cognitive and cognitive abilities compared to the median. These findings are consistent with a longer-term impact, measured in the Primary School Certificate (PSC) examination, where exam take-up rates have risen among initially less-able students.[9] Moreover, the PSC math grades are higher among treated school graduates than among similar students from control schools in terms of initial characteristics and/or likelihood of taking the exam. As an unintended impact, we also find that the intervention significantly improves teachers' ability to assess student performance, which suggests that accessing students' daily progress records has the potential to improve teachers' quality.

The remainder of this paper is organized as follows. In Section 2, we outline our experimental design, including the setting and intervention, followed by a description of the data and baseline test results. Section 3 gives the econometric evaluation framework, followed by empirical results. Section 4 addresses the comparison of benefits and costs of this intervention, and Section 5 concludes the paper.

II. Experiment Design, Data, and Balancing Test

A. Setting: BRAC Primary School

Primarily, BPS targets children from disadvantaged social backgrounds who could not get into formal schooling at the right age or have dropped out of the system. The economic eligibility criteria states that "children of poor households having less than 50 decimals of land and at least one member of the household has worked for wage for at least 100 days" and living within a two-kilometer radius of the school are admitted in BPS (Afroze, 2012). BPS basically covers the same standard curriculum as public schools. Up to grade three, BRAC develops textbooks and other materials, but government textbooks are used in grades four and five.

Although the BPS and government primary schools teach the same competency-based curriculum, there are some basic differences between them. Unlike the

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

[9]Every December, the Ministry of Primary and Mass Education conducts the PSC, and those who wish to advance to junior high school need to pass the exam. BPS is an accelerated program, so students thereof can take the PSC after completing the fourth grade if they are willing.

standard primary school system of five years, BPS offers an accelerated (four-year) program to bring these children back to the formal education track (Asadullah, 2016). In particular, BPS teachers cope with 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 six years for entry into primary education); the schools operate under a rather flexible time schedule for three hours a day, six days a week, with fewer holidays than government schools have, which results in higher contact hours per primary cycle than government primary schools have on average; the average class size in BPS (25-30 students) is about half that of government primary schools.

BPSs are essentially one classroom/one teacher schools, whereby a teacher teaches all subjects to the same cohort. The pedagogical approach is, however, influenced by traditional methods such as group lecture followed by assignments. Students are required to pass the grade five terminal examination set by the government, which also suggests that BPS teaches learners the same skills that are taught government schools, whereby teaching to the test potentially affects students' learning.

Thus, in this context, the Kumon intervention is aimed to promote self-learning by facilitating each student in studying at the right level and learning to set goals and take challenges to the next level. Given the unique setting of this non-formal education, such as the low-cost platform and smaller class size, BPS has the potential to scale up this intervention to improve primary education in Bangladesh through developing students' cognitive and non-cognitive abilities.

B. Intervention: The Kumon Method of Learning

The Kumon method of learning has been introduced in selected BPSs among third- and fourth-grade students as a supplementary module in mathematics. Kumon aims to enable students to develop advanced academic and self-learning abilities by ensuring that they always study at a level that is just right for each student. Students are assigned to an initial level based on their individual performance on a diagnostic test (DT) provided by the Kumon Institute of Education Co., Ltd., not on the basis of their school grade or age. The Kumon method is uniquely designed to set the initial level slightly lower than the student's concurrent maximum capacity in order to: i) ensure full understanding of the basic concepts as a firm building block for cognitive ability development; and ii) stimulate students' motivation to continue studying, which also works for the development of their non-cognitive abilities such as self-esteem and sense of competence. The Kumon mathematics program is divided into 20 levels (from Level 6A to Level O), and five elective levels, comprising a total of 4,420 double-sided worksheets. All of these worksheets are carefully designed, starting from simple counting to advanced mathematics, with the level of difficulty increasing in small steps.[10]

[10]Appendix A explains the details of the worksheets designed by Kumon, using a couple of worksheet examples. The final level of the material covers the high school graduation level.

Worksheets contain example questions with hints that help students to acquire step-by-step problem-solving skills by themselves. Kumon instructors do not provide lectures; they simply observe students' progress. They adjust the level of worksheets if students are stuck on the same worksheet or cannot find the right answer after many attempts. As a result, students can absorb material beyond their school grade level through self-learning and advance to high school-level material at an early age. Importantly, slower learners can spend more time on the basics without being rushed to move on to advanced-level materials beyond their level of understanding.

Another feature of Kumon is a tracking system for each student's progress and achievements using personalized record books. Kumon instructors do not teach in the class and, hence, do not need extensive prior experience to conduct daily quizzes to monitor each student's understanding and progress. This is because Kumon worksheets are laid out in small steps to enable students to self-learn, and there is a set standard time to solve each worksheet, which allows teachers to determine which level students can advance to the next level or should repeat a level. Having detailed progress reports on the worksheets allows instructors to obtain more objective information about their students' abilities and understanding of the mathematics involved.

C. Experimental Design

To identify the causal effects of Kumon on young students' cognitive and noncognitive abilities, we design and conduct an RCT study. For our purpose, we need a design that allows us to have adequate statistical power to detect a minimum effect of at least a 0.4 standard deviation, consistent with the effect size of education intervention elsewhere.[11] 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. These result in a sample of approximately 26 clusters with a statistical power of 0.80. To ensure that we do not lose statistical power due to attrition or other factors, we choose a cluster size of 34, with an average of 30 students per cluster (the average class size of BPS), giving us a sample of approximately 1,000 students.

We randomly select 34 BPSs comprising third- and fourth-graders from the 179 BPSs in Dhaka and its surrounding areas, with 17 schools receiving Kumon materials and 17 schools not receiving these materials so that they can serve as treatment and control schools, respectively.[12] The resulting sample breakdown

[11] Considering the results from some studies of high-impact education interventions that are teaching at the right level, such as Lakshminarayana et al. (2013) and Duflo, Dupas and Kremer (2011), we hypothesize a minimum detectible effect of 0.40 on cognitive ability for high policy impact.

[12] A stratified randomization at the school-branch level might have been more suitable in this situation; however, following a concern related to implementation challenges, we employ the method of randomization without stratification. To address concerns about potential spurious correlations between intervention and student outcomes arising from the unobserved heterogeneity across school-branches, we specify alternative models to conduct robustness checks. These are discussed in detail in Section 3.

by grade is as follows: 19 (out of 48 schools) for the third grade and 15 (out of 131 schools) for the fourth grade.[13] In these schools, we select only one of the two class shifts (either morning or afternoon), with an average class size of 30 students. The intervention consists of a 30-minute session on Kumon study prior to the beginning of their regular lessons. Thus, during the study periods, students in the treatment schools come to school earlier than the usual school hours.[14] BPS usually follows flexible hours and runs for six days a week except on public holidays and teacher training days. Our intervention lasted for eight months, from August 2015 to April 2016.

For the treatment schools, the Kumon Institute of Education Co., Ltd has provided an intervention package consisting of a mathematics materials set and an instructor manual with sheets for the BRAC teacher.[15] The full materials set consists of i) mathematics worksheets with questions at various difficulty levels; and ii) a grading notebook to record everyday progress, including the level of worksheet that a student works on, the number of repetitions required before achieving a full score on the worksheet, and the number of worksheets that students finally complete.[16]

During the administration of Kumon program, the BPS teachers do not provide lectures; they simply observe students' progress. They only intervene when students are stuck on the same worksheet or cannot find the right answer after many attempts. They adjust the level of worksheets in such cases. The BPS teachers also provide guidance when advanced students proceed to entirely new materials beyond the regular curriculum. The marking assistants help the teachers with grading and recording the worksheets. Until the session ends, students either move on to a new worksheet once they have achieved a full score on the previous one or continue to try and correct wrong answers until they have achieved a full score within the designated timeframe.

D. Data Description

We construct cognitive ability measures both at the baseline and endline based on two different mathematics test scores for both the treatment and control school students. These mathematics tests are developed by the Kumon Institute of

[13]The treatment schools do not overlap in terms of grade. In other words, in the treatment schools, Kumon intervention is applied to either the third or fourth grades.

[14]For practical purposes, our intervention departs from a standard Kumon center in two ways. First, students remain in the same classroom in which their regular BPS classes are held, while Kumon centers are normally outside school premises. Second, students are not given homework, unlike the standard practice in Kumon.

[15]BRAC field staff has been assigned to assist and follow up on BPS teachers. Three days of preparatory training for BPS teachers and field staff have been held prior to launching the program to familiarize teachers with the concepts and procedures of the learning method. In addition, three follow-up training sessions have been held during the implementation period. Two marking assistants have been provided for each class to support the grading and recording of worksheets during the Kumon session. BPS teachers monitor students and determine which level of worksheets that students work on.

[16]All the materials, including numbers, have been provided in the Bengali language, which is the medium of instruction for BPS teachers and students.

Education Co., Ltd. and are known as the Diagnostic Test (DT) and Proficiency Test of Self Learning (PTSII).[17]

The DT measures cognitive math abilities, whereby we retain records of both the score and the time taken to complete the test. The DT used for this study requires students to answer 70 questions within a maximum of 10 minutes. Hence, for the DT, we calculate test scores per minute (DT Score per min) to determine students' cognitive ability.

The PTSII has two sections: the first part consists of a total of 348 math questions within six categories measuring different dimensions of math problem-solving skills, whereby the aggregate score defines students' cognitive ability (PTSII-C). The second section consists of 27 questions, whereby the aggregate score captures students' non-cognitive ability (PTSII-NC) (See Appendix A). Among the 27 questions, 10 are consistent with the Children's Perceived Competence Scale (CPCS) (Sakurai and Matsui, 1992; Harter, 1979), and 8 are consistent with the Rosenberg Self-Esteem Scale (RSES) (Rosenberg, 1965). As non-cognitive ability measures, we employ the aggregated PTSII-NC index as well as the CPCS and RSES indices.

To assess the long-term impact of the intervention, we also collect students' results from the PSC examination, which is a nationally administered primary education completion test by the Ministry of Primary and Mass Education. Those who wish to continue to further education need to pass the exam, and based on the exam results, letter grades from A+ to A, A-, B, C, D, and F are assigned.[18] The subjects include math and English in addition to other subjects, but we focus on the math PSC results, given that our intervention is related to math problem-solving skills. Grade-four students had a chance to take the PSC exam about 8 months after the end of the intervention (December, 2016), while grade-three students took it about 20 months after (December, 2017).[19]

We also conduct a teacher survey, as well as a parent/guardian survey. The former data are employed for the analysis of teachers' assessment ability of student performance while the latter are used for the baseline balancing test to address the comparability of treatment and control school students in terms of household characteristics.

The sample attrition rates in our study between the baseline and endline are,

[17]Table B1 in Appendix B shows the list of data sets. Table B2 in Appendix B presents the descriptive statistics of major learning outcomes such as unconditional means of DT score per minute and PTSII-C score as well as non-cognitive test scores (RSES and CPCS consistent non-cognitive scores) of the control and treatment groups with the difference between the two groups at the baseline and endline. See also Appendix B regarding how the tests and survey results have been merged, as well as information on the unbalanced sample.

[18]The letter grades are assigned based on the exam scores: if the score is in the range of 80 to 100, the letter grade is an A+; if 70 to 79, it is an A; if 60 to 69, it is an A-; if 50 to 59, it is a B; if 40 to 49, it is a C; if 33 to 39, it is a D; and if below 33, it is an F. ⟨ `http://www.educationboard.gov.bd/computer/grading_system.php` ⟩

[19]Generally, this exam is administered at the end of the fifth grade as a primary school terminal examination. However, BPS adopts an accelerated curriculum that finishes at fourth grade, and the students are allowed to take the PSC at the end of the fourth grade.

on average, 11.3 percent in the treatment schools and 15.6 percent in the control schools. However, there is no systematic correlation between the attrition and observed characteristics.[20]

E. Balancing Test Results

We perform the baseline balance tests by comparing the main outcome variables of interest between the treatment and control group students: DT scores per minute, PTSII-C scores, and variables measuring non-cognitive abilities (PTSII-NC, RSES, CPCS) are all standardized to have a mean of zero and a standard deviation of one. Table 1 shows the baseline results of regressing each pre-intervention outcome on the treatment dummy, conditional on child and household characteristics, branch dummies, and test quality-adjustment variables. The test quality-adjustment variables include three dummy variables for time mismanagement (schools failing to restrict the test time or failing to assign a separate time limit for PTSII-C and PTSII-NC tests), mismatching of test level, and suspicion of cheating (27 students were reported).[21] The baseline outcome variables are balanced.

The unconditional balancing test results are reported in Table B2 of Appendix B, which shows some significant differences in the outcome variables. Firstly, we observe a faster DT completion time among the treatment group than in the control group students at the baseline, which results in a higher DT score per min. We can reasonably attribute the faster DT completion time among the treatment group to the higher proportion of cheating identified on the DT in the treatment group than in the control group. Secondly, we observe that PTSII-C scores are higher for control group students. This result can be attributed to the mismanagement of the PTSII test time: this has occurred more among the control group than in the treatment group in both directions, whereby students were given either a shorter time than the set time (6% treatment vs 12% control schools) or unlimited time (6% treatment vs 13% control schools). The shorter time might have affected the PTSII-NC (the latter half of PTSII) and most likely resulted in missing answers at the baseline as we observe. The unlimited time, on the other hand, could lead students to score higher in PTSII-C at the baseline. For these reasons, we could observe a higher average PTSII-C score among the

[20]See Appendix B, which shows the characteristics of dropouts and the sample used in the analysis. To calculate attrition rates, we consider a student to be a dropout if he/she did not take either the DT or the PTSII at the endline. In treatment schools, 57 out of 478 students, and in control schools, 82 out of 526 students did not take either the DT or the PTSII at the endline for various reasons (e.g., dropout, absence on exam days, switch of schools, etc.).

[21]The details of the test quality-adjustment variables are given as follows. Time mismanagement dummies: school nos. 8, 26, and 31 did not comply with the time restriction, and school nos. 9, 24, and 25 failed to allocate a separate time for the PTSII-C and PTSII-NC; and mismatching of test level dummy: Out of two levels of the DT, school nos. 14, 18, 19, and 20 in grade three took the wrong DT (Level P3 instead of P1). The DT (Level P2) used for the analysis is not directly affected, but we include a dummy variable to control for any possible indirect effect. Suspicion of cheating dummy: Based on teachers' reports and Kumon's assessment of an observed gap between DT results and the starting level of worksheets, 27 cases are reported.

control group. We believe these baseline imbalances do not weaken our main conclusion in the paper because both the included cheating dummies and time-mismanagement dummies can absorb for potential bias in all our estimations.

III. Empirical Specification and Results

A. Students' Learning Outcomes

Econometric specification

We employ the canonical difference-in-differences model to estimate the impact of the Kumon intervention on our measures of cognitive as well as non-cognitive abilities of student i at time t, Y_{it}: $Y_{it} = \alpha_0 + \alpha_1 T_t + \gamma d_i + \delta T_t \cdot d_i + u_i + \varepsilon_{it}$, where the Kumon intervention is specified by an indicator variable, d, taking 1 for the treatment group and 0 for the control group; T is a time dummy; and u and ε are student fixed effects and the error term, respectively. The average treatment effects on the treated can be captured by estimated δ. For the estimation, we take the first difference of the original level equation, whereby the dependent variable captures improvements in cognitive or non-cognitive outcomes:

$$\Delta Y_{it} = \alpha_1 + \delta d_i + \Delta \varepsilon_{it}, \tag{1}$$

where Δ is a first-difference operator. 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).[22]

To investigate heterogeneous treatment effects, we estimate the equation (1) for four different sub-samples: i) high-initial cognitive ability and non-cognitive ability students (high-high type); ii) high-initial cognitive ability and low-initial non-cognitive ability students (high-low type); iii) low-initial cognitive ability and high-initial non-cognitive ability students (low-high type); iv) low-initial cognitive ability and low-initial non-cognitive ability students (low-low type). The cut-off points for high and low are the median value of respective outcome measures at the baseline. The parameters of interests are δ for different initial ability types.

[22] 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.).

Table 1— Students' Cognitive and Non-cognitive Abilities: Baseline Balancing Test Results

Dependent Variables	DT Score per min[a] (1)	DT Score (2)	DT Time (3)	PTSII-C Score[b] (4)	RSES[c] (5)	CPCS[c] (6)
Treatment	0.018	0.014	-0.015	-0.195	0.027	0.153
	(0.141)	(0.157)	(0.075)	(0.132)	(0.119)	(0.116)
Constant	-0.026	-0.010	0.034	-0.023	-0.001	-0.082*
	(0.121)	(0.129)	(0.046)	(0.117)	(0.096)	(0.085)
Number of Observations	968	968	968	1,004	1,004	1,004
R-squared	0.023	0.010	0.150	0.327	0.008	0.010

Notes: Asymptotic standard errors are shown in parentheses and are clustered at the school level (34 clusters). The asterisks reflect the significance levels obtained by a clustered wild bootstrap-t procedure; ***, **, and * denote the 1 percent, 5 percent, and 10 percent levels, respectively. All regressions are controlled for branch-fixed effects. We also include dummy variables for cases of suspected cheating, misguidance of test time, and mismatching of test level. Control variables: number of members in the household, number of adults in the household, number of members in the household who have completed primary education, number of males in the household, availability of electricity, availability of gas connection, source of water, house ownership, dummy for missing variables, and interaction terms between dummy for missing variables and other covariates.
[a]: DT Score per Minute stands for Diagnostic (Math) Test score per minute: 70 questions must be solved correctly in 10 minutes.
[b]: PTSII-C Score stands for the Proficiency Test of Self Learning first half score, which consists of 348 math questions.
[c]: Among the second half of the Proficiency Test of Self Learning, consisting of 27 survey questions prepared by Kumon, 10 are consistent with the Children's Perceived Competence Scale (CPCS Non-cognitive Score) and 8 with the Rosenberg Self-Esteem Scale (RSES Non-cognitive Score). For survey questions related to each Non-cognitive Score, see Appendix C. The responses are recorded on a four-point scale: 1=Strongly Agree, 2=Somewhat Agree, 3=Somewhat Disagree, 4=Strongly Disagree. Both cognitive and non-cognitive test scores are standardized and used in the regression analysis.

Results of cognitive and non-cognitive abilities

The first four columns of Table 2 report the results of the estimating equation 1, using cognitive outcomes that are standardized, so that the magnitudes of the impacts are reported in terms of their standard deviations. As shown, we find significant improvements in the cognitive outcomes measured by DT score per minute and PTSII-C scores. The magnitude of the impact is enormous: a 2.177 standard deviation in terms of DT scores per minute. While this effect size may seem surprisingly high compared to the effect size of education interventions elsewhere, it should be noted here that effect size on DT score per minute is coming through substantial reduction in test completion time measured as DT time (-2.274 s.d.). However, the effect size of the DT score (0.505 s.d.), i.e., improvement in raw test score, is consistent with previous findings in literature. Unlike previous studies that have used test scores to determine cognitive ability, we use test score per minute (DT score per minute), as our intervention is designed to increase students' ability to solve math problems in a time-efficient manner, which is important in pursuing higher education with more complex materials. We also employ an alternative measure of cognitive ability, PTSII-C, to estimate equation (1). As we can see, the estimated effect size using PTSII-C is a 1.198 standard deviation.[23] While part of the improvement could result from the fact

[23]These findings are robust in ANCOVA specifications, which are supposed to be less sensitive to natural within-person variation, unlike DID, in the baseline and endline variables (McKenzie, 2012). See Appendix D.

Table 2— Impact of Kumon on Students' Cognitive Abilities: DID Estimates

Dependent Variables	DT Score per min[a] (1)	DT Score (2)	DT Time (3)	PTSII-C Score[b] (4)	RSES (5)	CPCS (6)
Treatment	2.177***	0.505**	-2.274***	1.198***	-0.043	-0.113
	(0.531)	(0.208)	(0.512)	(0.209)	(0.177)	(0.180)
Constant	1.109	0.583*	-0.955	0.565	0.017	0.102
	(0.320)	(0.167)	(0.339)	(0.140)	(0.091)	(0.101)
Num of Obs.	799	799	799	787	696	696
R-squared	0.208	0.064	0.217	0.423	0.026	0.020

Notes: Asymptotic standard errors are shown in parentheses and are clustered at the school level (34 clusters). The asterisks reflect the significance levels obtained by a clustered wild bootstrap-t procedure; ***, **, and * denote the 1 percent, 5 percent, and 10 percent levels, respectively.
[a]: DT Score per Minute stands for Diagnostic (Math) Test Score per minute: 70 questions must be solved correctly in 10 minutes. Both cognitive and non-cognitive test scores are standardized and used in the regression analysis.
[b]: PTSII-C Score stands for the Proficiency Test of Self Learning first half score, which consists of 348 math questions.

that the treated group has become comfortable with taking paper-based math quizzes, the sizable impact on cognitive ability measurements suggest that the self-learning approach had substantially enhanced their numeracy skills (particularly their arithmetic skills). When we examine DT score and DT time separately, it emerges that the large impacts on DT score per minute largely result from the improved math-problem-solving speed measured by DT time. Moreover, the magnitude of the effect on PTSII-C scores is in line with that found by past studies that have focused on teaching at the right level (Lakshminarayana et al., 2013; Duflo, Dupas and Kremer, 2011). In contrast, regarding the non-cognitive outcomes reported in the last two columns of Table 2, the homogeneous treatment effects estimates are insignificant. Several hypotheses are being tested at the same time: six in Table 2. We have adjusted p-values for multiple testing by the Romano-Wolf procedure (Romano and Wolf, 2005), finding qualitatively the same results of statistical inferences. Also, we confirm these qualitative results reported in Table 2 using endline data only (Table 3). Furthermore, an analysis of covariance (ANCOVA) specification with baseline outcomes as covariates gives the same results qualitatively (Appendix D).

The heterogeneous treatment effects are reported in Table 4. We find positive and significant coefficients on cognitive outcomes for all four initial ability types. The magnitudes on DT score per minute are largest for the students with high-initial cognitive and non-cognitive abilities (high-high type), while they are smallest for the students with low-initial abilities in both measures (low-low type). Regarding the non-cognitive outcomes, however, we find a catching-up effect: students with initially low cognitive and non-cognitive abilities (low-low type) show a positive and significant treatment effect on the change in non-cognitive scores (RSES) while others do not show significant effects in non-cognitive scores. These results suggest a building block hypothesis of non-cognitive ability: the Kumon intervention first improves non-cognitive ability of those who are initially lag-

Table 3— Endline Cross Section

Dependent Variables	DT Score per min[a] (1)	DT Score (2)	DT Time (3)	PTSII-C Score[b] (4)	RSES [c] (5)	CPCS [c] (6)
Treatment	2.146***	0.463***	-2.300***	0.947***	0.003	0.100
	(0.511)	(0.125)	(0.504)	(0.214)	(0.150)	(0.144)
Mismatch of test level	0.967	-0.258***	-1.369	-0.475**	0.290	0.226
	(0.503)	(0.141)	(0.442)	(0.370)	(0.263)	(0.238)
Cheating	-1.267***	-0.097**	1.225***	-0.058**	-0.178	-0.191
	(0.484)	(0.171)	(0.380)	(0.222)	(0.289)	(0.286)
Mismanagement of time (Shorter)	0.185	-0.609*	-1.186**	-0.672**	-0.322	-0.476
	(0.197)	(0.158)	(0.562)	(0.199)	(0.198)	(0.237)
Mismanagement of time (Unlimited)	0.169	-0.063	-0.449	-0.072	0.158	0.074
	(0.198)	(0.119)	(0.261)	(0.509)	(0.239)	(0.218)
Grade 3 dummy	-0.710	0.072	0.868	0.713***	-0.020	-0.003
	(0.431)	(0.132)	(0.448)	(0.194)	(0.146)	(0.136)
Constant	1.168	0.655**	-0.976	0.548	-0.029	-0.052
	(0.293)	(0.129)	(0.334)	(0.140)	(0.107)	(0.098)
Observations	811	811	811	837	832	832
R-squared	0.234	0.129	0.321	0.220	0.021	0.031

Notes: Asymptotic standard errors are shown in parentheses and are clustered at the school level (34 clusters). The asterisks reflect the significance levels obtained by a clustered wild bootstrap-t procedure; ***, **, and * denote the 1 percent, 5 percent, and 10 percent levels, respectively.

[a]: DT Score per Minute stands for the Diagnostic (Math) Test Score per minute: 70 questions must be solved correctly in 10 minutes.

[b]: PTSII-C Score stands for the Proficiency Test of Self Learning first half score, which consists of 348 math questions.

[c]: Among the second half of the Proficiency Test of Self Learning, consisting of 27 survey questions prepared by Kumon, 10 are consistent with the Children's Perceived Competence Scale (CPCS Non-cognitive Score) and 8 with the Rosenberg Self-Esteem Scale (RSES Non-cognitive Score). For survey questions related to each Non-cognitive Score, see Appendix C. The responses are recorded on a four-point scale: 1=Strongly Agree, 2=Somewhat Agree, 3=Somewhat Disagree, 4=Strongly Disagree. Both cognitive and non-cognitive test scores are standardized and used in the regression analysis.]

ging in both cognitive and non-cognitive abilities; then, in turn, it improves the cognitive ability of those with sufficiently improved non-cognitive ability.

Since students in the treatment schools have studied Kumon materials for an additional 30 minutes per day, one might argue that the impact estimates we present here can also be due to longer session times in schools and not merely due to the Kumon intervention. To test such a possibility, we exploit the fact that some treatment schools have conducted Kumon sessions for at least five minutes longer. Using these time variations in the Kumon sessions, we examine the impact of longer study time of Kumon (Table 5).[24] Insignificant coefficients on the cross-term between the treatment and longer-session dummy suggest that overall outcomes are not systematically affected by a longer school session. Therefore, we believe that the impact observed in this study can be attributed to the Kumon method of learning.

Long-term impact

To assess the long-term impact of the intervention, we collect additional information regarding national examination achievements after 8 months and 20 months of intervention, respectively, for the grade-four and grade-three students in our sample. Specifically, we gather the PSC examination results as well as the reasons for dropouts, if any. From our sample, 43 (37) and 54 (53) percent of grade-three and grade-four students, respectively, from the treatment (control) schools took the exam in November-December 2016 and 2017, respectively.[25] Since the proportion of students who took the exam is higher in the treatment schools than in the control schools, we need to address the potential selection bias when comparing the PSC outcomes of the two groups. Indeed, among those who took the PSC exam, the average initial DT score of the treatment school students is significantly lower than that of the control school students.[26]

To address potential selection bias, we employ alternative specifications in estimating the impact of Kumon on PSC exam participation and results in PSC

[24]There is also evidence that extra hours of tutoring do not have a significant impact on test scores of NGO primary school students in Bangladesh, although they do reduce dropout rates (Ruthbah et al., 2016).

[25]We collected students' PSC registration IDs from the BPS branch offices and teachers of the schools. Then we obtained their PSC results from the government websites based on the IDs. We also collected information from the schools about dropouts from the PSC (non-takers). As described, the PSC take-up rate is relatively higher among treatment school students. The primary reason for not taking the primary terminal examination was family relocation (79 percent), while other reasons included dropouts due to 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 no longer studying due to other reasons (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 when a child's family relocates from the area during this period, it is highly likely that they will 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 or about dropouts.

[26]The mean DT score of PSC takers from the treatment schools is -0.021, while that of control school is 0.266, which is significantly different by 0.287 (0.092) at 1 percent significance level. Similarly, the mean of PTS-C scores among PSC takes at the treatment schools is -0.100, while that of the control schools is 0.328, which is significantly different by 0.428 (0.087) at 1 percent significance level.

Table 4— Heterogenous Impact of Kumon on Students' Cognitive Abilities: DID Estimates

	Initial RSES Score					Initial CPCS Score				
Dependent Variables	DT Score per min[a]	DT Score	DT Time	PTSII-C Score[b]	RSES	DT Score per min[a]	DT Score	DT Time	PTSII-C Score[b]	CPCS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: High Initial Cognitive High Initial Non-cognitive										
Treatment	3.219***	0.453**	-3.396***	1.254***	0.044	3.284***	0.438**	-3.339***	1.240***	0.075
	(0.764)	(0.182)	(0.637)	(0.254)	(0.203)	(0.864)	(0.174)	(0.683)	(0.267)	(0.196)
Constant	0.332	-0.116**	-0.721	0.217	-0.693***	0.634	-0.028	-1.089	0.148	-0.704***
	(0.311)	(0.125)	(0.306)	(0.232)	(0.154)	(0.389)	(0.136)	(0.432)	(0.223)	(0.146)
Num of Obs.	186	186	186	189	189	188	188	188	188	188
R-squared	0.337	0.160	0.339	0.474	0.037	0.319	0.137	0.310	0.489	0.060
Panel B: High Initial Cognitive Low Initial Non-cognitive										
Treatment	3.092***	0.286	-3.720***	1.248***	-0.190	2.843***	0.265	-3.603***	1.306***	-0.209
	(0.897)	(0.188)	(0.689)	(0.218)	(0.234)	(0.788)	(0.223)	(0.696)	(0.226)	(0.289)
Constant	0.840**	0.170	-1.065**	0.087	0.588***	0.523*	0.074	-0.663**	0.152	0.663***
	(0.370)	(0.135)	(0.457)	(0.149)	(0.174)	(0.258)	(0.105)	(0.306)	(0.162)	(0.191)
Num of Obs.	169	169	169	180	180	167	167	167	181	181
R-squared	0.331	0.108	0.392	0.475	0.073	0.312	0.109	0.389	0.457	0.032
Panel C: Low Initial Cognitive High Initial Non-cognitive										
Treatment	2.480**	0.628**	-2.504***	1.538***	0.020	2.290**	0.549**	-2.323**	1.354***	0.212
	(0.939)	(0.243)	(0.801)	(0.292)	(0.231)	(0.973)	(0.255)	(0.862)	(0.322)	(0.286)
Constant	1.656***	1.033***	-0.995**	0.648***	-0.674***	1.829	1.229***	-0.935	0.818***	-0.656***
	(0.433)	(0.216)	(0.410)	(0.211)	(0.146)	(0.453)	(0.169)	(0.430)	(0.248)	(0.221)
Num of Obs.	150	150	150	150	150	141	141	141	150	150
R-squared	0.201	0.109	0.258	0.352	0.033	0.166	0.076	0.205	0.292	0.050
Panel D: Low Initial Cognitive Low Initial Non-cognitive										
Treatment	1.158***	0.463	-1.532***	1.134***	0.412**	1.434***	0.589**	-1.760***	1.248***	0.230
	(0.377)	(0.253)	(0.499)	(0.226)	(0.164)	(0.334)	(0.206)	(0.483)	(0.252)	(0.244)
Constant	1.904***	1.503***	-0.846	0.983**	0.537***	1.704*	1.311***	-0.863	0.882*	0.589***
	(0.295)	(0.187)	(0.371)	(0.173)	(0.0988)	(0.282)	(0.190)	(0.423)	(0.185)	(0.122)
Num of Obs.	171	171	171	177	177	180	180	180	177	177
R-squared	0.167	0.057	0.245	0.312	0.051	0.226	0.111	0.289	0.354	0.028

Notes: Asymptotic standard errors are shown in parentheses and are clustered at the school level (34 clusters). The asterisks reflect the significance levels obtained by a clustered wild bootstrap-t procedure; ***, **, and * denote the 1 percent, 5 percent, and 10 percent levels, respectively.

[a]: DT Score per Minute stands for Diagnostic (Math) Test Score per minute: 70 questions must be solved correctly in 10 minutes. Both cognitive and non-cognitive test scores are standardized and used in the regression analysis.

[b]: PTSII-C Score stands for the Proficiency Test of Self Learning first half score, which consists of 348 math questions.

Table 5— Impact of Kumon on Students' Cognitive and Non-cognitive Abilities: Estimates Controlling for Longer Sessions

Dependent Variables	DT Score per min[a] (1)	DT Score (2)	DT Time (3)	PTSII-C Score[b] (4)	RSES [c] (5)	CPCS [c] (6)
Treatment	2.368***	0.353	-2.646***	1.188***	-0.118	-0.176
	(0.680)	(0.225)	(0.564)	(0.248)	(0.184)	(0.191)
Treatment x Longer session	-0.578	0.459*	1.124*	0.027	0.286	0.241
	(0.541)	(0.229)	(0.586)	(0.234)	(0.271)	(0.268)
Constant	1.103***	0.587***	-0.944***	0.566***	0.014	0.099
	(0.314)	(0.162)	(0.322)	(0.140)	(0.088)	(0.098)
Num of Obs.	799	799	799	787	696	696
R-squared	0.214	0.082	0.242	0.423	0.031	0.024

Notes: Asymptotic standard errors are shown in parentheses and are clustered at the school level (34 clusters). The asterisks reflect the significance levels obtained by a clustered wild bootstrap-t procedure; ***, **, and * denote the 1 percent, 5 percent, and 10 percent levels, respectively.
[a]: DT Score per Minute stands for the Diagnostic (Math) Test Score per minute: 70 questions must be solved correctly in 10 minutes.
[b]: PTSII-C Score stands for the Proficiency Test of Self Learning first half score, which consists of 348 math questions.
[c]: Among the second half of the Proficiency Test of Self Learning, consisting of 27 survey questions prepared by Kumon, 10 are consistent with the Children's Perceived Competence Scale (CPCS Non-cognitive Score) and 8 with the Rosenberg Self-Esteem Scale (RSES Non-cognitive Score). For survey questions related to each Non-cognitive Score, see Appendix C. The responses are recorded on a four-point scale: 1=Strongly Agree, 2=Somewhat Agree, 3=Somewhat Disagree, 4=Strongly Disagree. Both cognitive and non-cognitive test scores are standardized and used in the regression analysis.]

Table 6— Impact of Kumon on Primary School Certificate (PSC) Examination Math Results: DID, DID with Heckman's Two-step Estimates, Propensity Score Matching, and Inverse Probability Weighting

Method	DID (1)	DID-Heckman (2)	PSM (3)	DID-IPW (4)
Treatment	0.185	0.270*	0.222*	0.252*
	(0.358)	(0.145)	(0.119)	(0.153)
Constant	-0.234	-1.481**		
	(0.336)	(0.605)		
Num of Obs.	461	459	456	454
R-squared	0.006			

Notes: The dependent variable is a first-difference of Math PSC score. Asymptotic standard errors are shown in parentheses and are clustered at the school level (34 clusters). The asterisks reflect the significance levels obtained by a clustered wild bootstrap-t procedure; ***, **, and * denote the 1 percent, 5 percent, and 10 percent levels, respectively.

Mathematics. As shown in Table 6, we employ simple difference-in-differences (DID), DID with Heckman's sample-selection correction approach (DID-Heckman), Propensity Score Matching (PSM), and DID with Inverse Probability Weighting

(DID-IPW). The latter three are for taking into account a self-selection bias in taking the PSC examination.[27] For the DID with Heckman's two-step, the first-stage equation is to regress the choice of taking the exam on student age, gender, grade dummy, and the treatment dummy. The impact of the intervention is then examined, eliminating selection bias. For PSM and IPW, we match the sample based on pre-treatment student characteristics (i.e., student age, gender, and grade dummy). The results suggest that a significantly higher percentage of students from the treatment schools received slightly better grades (above B and C grades) than those from control schools (Table 6).[28] There is no difference in the exam passing rate or likelihood of scoring more than A or A+. Overall, we find a modest long-term impact of the intervention measured by the national-level examination given outside the purview of our experiment while after the completion of the intervention.

B. Teacher Assessment Ability

In addition to student outcomes, we also examine the impact of intervention on teachers' ability to assess student performance. We hypothesize that teachers can potentially improve their own understanding and assessment of students' abilities, as the intervention will allow them to gain more information about students' skills from the daily progress records.

We collect each teacher's evaluation of individual students' performances. We then take the absolute distance between teachers' evaluations and observed cognitive outcomes (DT Score per min or PTSII-C score).[29] Using this outcome measure, we conduct the same DID specification as equation 1.

Our findings on the improvement in teachers' ability to assess student performance are reported in Table 7. As shown in this table, we find a significant improvement in teachers' ability to assess student performance in both types of tests (i.e., a negative sign indicates that the assessment scale is closer to the actual test score scale).

These positive impacts on the BPS teachers are unintended but not a surprise, given the nature of the intervention. The BPS teachers interact with the program to the extent that they ensure that students comply with the intervention, i.e., study at the right level. By observing the study behavior and daily progress, the teachers can gain a precise idea of each students ability. While it may suggest

[27]We use DID to control for time-invariant unobserved differences between the treatment and control school students that affect both the decision to take the PSC and the results themselves. DID with Heckman's sample-selection correction model is introduced to further utilize the non-linearity of the inverse-mills ratio, which is calculated based on the estimated probability of taking the PSC, to identify the Kumon effect on PSC results, controlling for the individual fixed effect and selection on taking the PSC. PSM compares students with similar observable characteristics before the intervention; thus, the selection on measured cognitive and non-cognitive abilities is controlled. Lastly, DID with Propensity Score Weighting is a combination of DID and PSM in essence.

[28]The PSC grading scale is shown at the following link: ⟨ `http://www.educationboard.gov.bd/computer/grading_system.php` ⟩

[29]The students' test scores are categorized into a 1-5 scale to match the teacher's evaluation score.

Table 7— Association between Teachers' Assessment and Student Performance

Difference between Teachers' Objective Evaluation and Students' Objective Performance	DT Score per min[a] (1)	PTSII-C Score[b] (2)
Treatment	-0.348***	-0.350**
	(0.126)	(0.132)
Constant	1.535***	1.535***
	(0.110)	(0.110)
Num of Obs.	1,416	1,416
R-squared	0.050	0.047

Notes: Asymptotic standard errors are shown in parentheses and are clustered at the school level (34 clusters). The asterisks reflect the significance levels obtained by a clustered wild bootstrap-t procedure; ***, **, and * denote the 1 percent, 5 percent, and 10 percent levels, respectively.

[a]: DT Score per Minute stands for the Diagnostic (Math) Test Score per minute: 70 questions must be solved correctly in 10 minutes.

[b]: PTSII-C Score stands for the Proficiency Test of Self Learning first half score, which consists of 348 math questions.

that teachers could have modified their teaching in program schools, we find no significant difference in teaching hours or home work load between treatment and control schools. We agree that better information about students' progress makes teachers in treatment schools more accurate in their assessment of students abilities. However, our intervention promoting individualized self-learning is different from the diagnostic feedback interventions of Muralidharan and Sundararaman (2010) and de Hoyos, Ganimian and Holland (2017), whereby baseline test results are provided to teachers/schools in order to test its impact on teaching as well as student learning. The Kumon learning approach has good potential for reducing teacher stereotyping of students by providing them with better information about their students and encouraging teaching to learning instead "teaching to the test."

IV. Comparing Costs and Benefits

Following Duflo (2001) and Heckman et al. (2010), we calculate the benefit-cost ratio (B-C ratio) and internal rate of return (IRR). Regarding benefits, we use our long-term impact estimate on math PSC scores (Table 6) and estimated wage returns to numeracy skills from Nordman, Sarr and Sharma (2015) that use the matched employer-employee data. The 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. The first estimate is taken from our results on the PSC exam, and we use the most conservative number (DID-Heckman estimates), 0.212, in Table 6. The wage returns to numeracy skills, 0.037, are taken from Table 3, column 8 of Nordman, Sarr and Sharma

(2015). The average annual earnings are calculated based on the average hourly wage in Table 2 (50.91), multiplied by 40 hours per week and 52 weeks. The life cycle profile of earnings is calculated based on the estimates of the returns to tenure and tenure squared terms in the same regression we use for returns to numeracy skills (0.037 and -0.00067).

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

To construct the B-C ratio chart, we assume that the benefit will last from 1 year to 44 years, considering working for a lifetime from age 16 to 59 and an annual discount rate of 5 percent following Duflo (2001). The dead-weight loss factor is unused because this program did not involve tax spending or revenue. Under the minimum (maximum) cost assumption, the benefit to cost ratio exceeds one when the benefits last for more than fifteen (more than eighteen) years, as shown in Figure 1 (Figure 2). It should be noted, however, 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 that the present values of benefit and cost equalize over a specified time-horizon, varying from 1 year to 44 years. The IRR becomes positive when workers continue working with benefits for more than ten (twelve) years with the minimum (maximum) cost (Figures 1 and 2).

V. Conclusions

In this paper, we have investigated the effectiveness of a novel individualized self-learning method in overcoming the issue of low-quality teaching and learning in a developing country. Specifically, we have designed and implemented a field experiment to test the effectiveness of the Kumon mathematics learning program on improving primary school students' cognitive and non-cognitive abilities in Bangladesh. As an effective program to strengthen cognitive and non-cognitive learning outcomes, Kumon is based on a just-right level of study so that students are provided with a suitable amount of mental stimulus to enhance their academic and self-learning abilities. As an overall impact, after eight months of intervention, we find significant and robust improvements in students' cognitive abilities measured by diagnostic test scores per minute and proficiency test scores. The magnitude of this impact ranges from a 0.505 to a 2.177 standard deviation where the upper and lower bounds are measured by diagnostic test scores per minute and time-unadjusted test scores, respectively. These impacts on cognitive ability are consistent with some existing interventions such as the 0.75 standard devia-

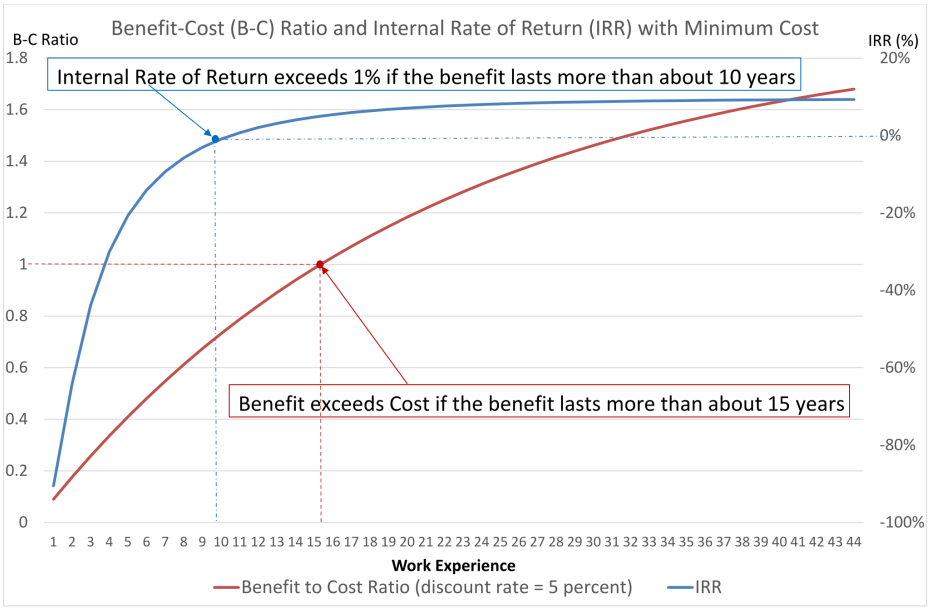


Figure 1. Benefit-Cost (B-C) Ratio and Internal Rate of Return (IRR) with Minimum Cost

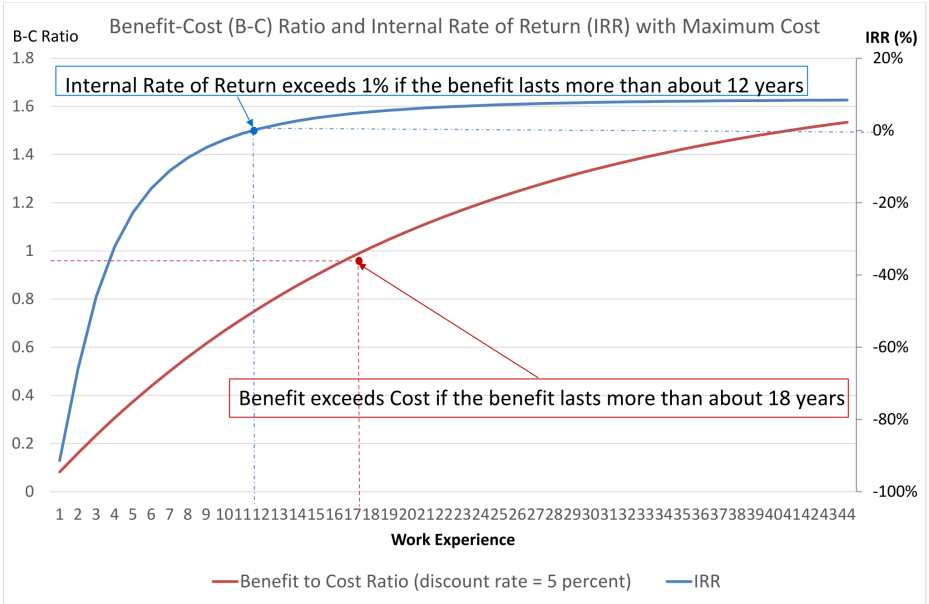


Figure 2. Benefit-Cost (B-C) Ratio and Internal Rate of Return (IRR) with Maximum Cost

tion impact of the supplementary remedial teaching provided by Indian NGOs to pupils in public primary schools (Lakshminarayana et al., 2013). Regarding non-cognitive abilities, we find catch-up effects among the initially low non-cognitive and low cognitive ability pupils. Furthermore, we have demonstrated the long-term impact of the intervention as measured by students' achievements on the national-level examination taken 8 and 20 months after the intervention. Lastly, we have found some positive impacts on BPS teachers' capacity to assess student performance. This latter finding implies that BPS teachers might have benefited from the Kumon intervention by gaining more objective information about students' skill levels.

The contributions of this paper are summarized as follows: By demonstrating the effectiveness of an innovative "self-learning at the right level" method on students' cognitive and non-cognitive abilities as well as long-term outcomes, we believe our study contributes substantially to the existing literature focusing on improving the quality of primary education in developing countries, in particular, the literature that examines the effectiveness of pedagogical interventions on student learning outcomes (Duflo, Dupas and Kremer, 2011; Banerjee et al., 2016, 2007; Muralidharan, Singh and Ganimian, 2019). Since the Kumon method of learning has already been extended globally, our results are potentially generalizable to similar socioeconomic and policy contexts. We have also provided a benefit-cost comparison of the intervention, showing that the benefits will outweigh the costs in future years if the effect (in terms of labor market outcome) lasts for ten to twelve years or more.

From the policy perspective, this study demonstrates that Kumon could be an effective complementary intervention for the existing lecture-style primary education for disadvantaged students such as dropouts from formal education and those with a low socioeconomic status. Moreover, unlike the existing successful computer-assisted learning programs, the Kumon method of learning is not constrained by inadequate electricity supplies we often encountered in developing countries.

Acknowledgement

Funding

This work was supported by a Grant-in-Aid for Scientific Research (S) from the Japan Society for the Promotion of Science (KAKENHI 26220502).

Conflict of interest

We are thankful to Esther Duflo, David Figlio, Deon Filmer, Dean Karlan, Halsey Rogers, Pascaline Dupas, and Paul Romer as well as the session participants at the American Economic Association Meeting 2018, World Bank Education GP BBL, the 2017 European, North American Summer, and Asian Meetings of the Econometric Society, the Australasian Development Economics Workshop

2017, the Midwest International Economic Development Conference 2017, the GRIPS-University of Tokyo Workshop 2017, Hitotsubashi University, the Kansai Labor Economics Workshop, and Hayami Conference 2016, for their useful comments. We are grateful to the authorities of BRAC, Kumon Institute of Education Co., Ltd., and Japan International Cooperation Agency (JICA) for their cooperation in implementing the study.

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, M 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, and Esther Duflo. 2006. "Addressing Absence." *Journal of Economic Perspectives*, 20(1): 117–132.

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.

Barrow, Lisa, Lisa Markman, and Cecilia Elena Rouse. 2009. "Technology's Edge: The Educational Benefits of Computer-aided Instruction." *American Economic Journal: Economic Policy*, 1(1): 52–74.

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.

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.

de Hoyos, Rafael, Alejandro J. Ganimian, and Peter A. Holland. 2017. "Teaching with the test: experimental evidence on diagnostic feedback and capacity building for public schools in Argentina." The World Bank.

Duckworth, Angela L., Christopher Peterson, Michael D. Matthews, and Dennis R. Kelly. 2007. "Grit: Perseverance and Passion for Long-Term Goals." *Journal of Personality and Social Psychology*, 92(6): 1087–1101.

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, and Michael Kremer. 2005. "Use of Randomization in the Evaluation of Development Effectiveness." In *Evaluating Development Effectiveness*. Vol. 7, , ed. George Keith Pitman, N. Ã. Osvaldo and Gregory K. Ingram, 205–231. New Brunwick, NJ.

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

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2007. "Using Randomization in Development Economics Research: A Toolkit." In *Handbook of Development Economics*. Vol. 4, , ed. T. Paul Schultz and John A. Strauss, 3895–3962. Elsevier.

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. 2002. "Schools and Skills in Developing Countries: Education Policies and Socioeconomic Outcomes." *Journal of Economic Literature*, 40(2): 436–482.

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., 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.

Jensen, Robert. 2010. "The (Perceived) Returns to Education and the Demand for Schooling." *Quarterly Journal of Economics*, 125(2): 515–548.

Kremer, Michael. 2003. "Randomized Evaluations of Educational Programs in Developing Countries: Some Lessons." *American Economic Review*, 93(2): 102–106.

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.

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.

Miguel, Edward, and Michael Kremer. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica*, 72(1): 159–217.

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.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2010. "The impact of diagnostic feedback to teachers on student learning: experimental evidence from India." *Economic Journal*, 120(546): F187–F203.

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.

Ruthbah, Ummul, Atonu Rabbani, Salim Hossain, and Golam Sarwar. 2016. "Do extra hours of tutoring payoff? Evaluation of a Community Education Programme in Bangladesh." *Journal of Development Effectiveness*, 8(2): 196–215.

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.

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

UNESCO. 2015. "Education for All Global Monitoring Report 2015/6 Education for all 2000–2015: Achievements and Challenges." Paris, France: UNESCO.

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

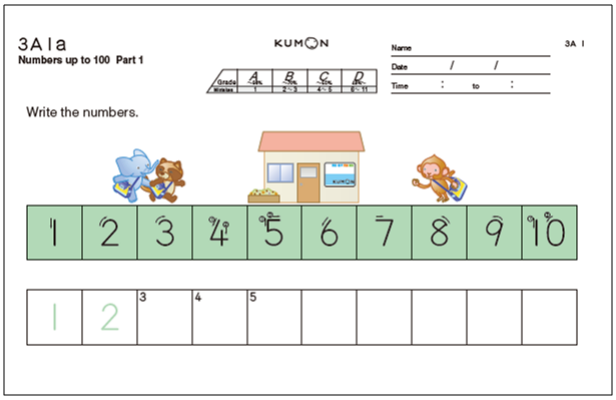
Watkins, Kevin. 2000. *The Oxfam Education Report.* Oxford: Oxfam International.

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

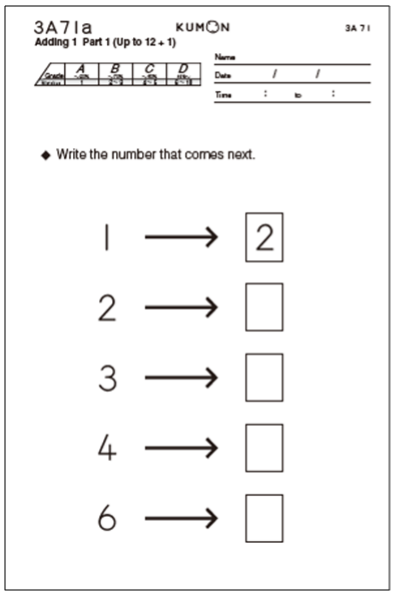
For Online Publication

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, which helps them to learn autonomously. For example, the first worksheet (3A1a) allows students to learn the order of numbers (up to 100, for example). Once students have mastered these worksheets without error within a targeted timeframe, they begin to learn the concept of addition (note: completion within a targeted time is a proxy for letting students 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 one 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 as well as the preceding worksheets make it possible for students to self-learn calculation skills and some of the math concepts behind them. Please note that these worksheets comprise the English versions thereof. In the case of the BRAC primary school trail, all materials were translated into Bengali, the local language that BRAC Primary School students regularly use in class.

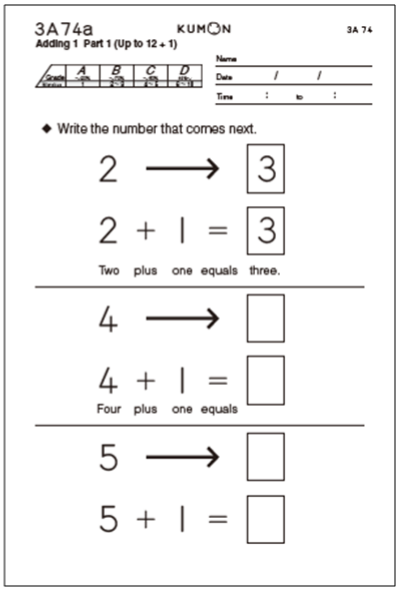


3A71a KUMON 3A 71
Adding 1 Part 1 (Up to 12 + 1)

Name
Date / /
Time : to :

◆ Write the number that comes next.

1 ⟶ 2
2 ⟶ ☐
3 ⟶ ☐
4 ⟶ ☐
6 ⟶ ☐



3A74a KUMON 3A 74
Adding 1 Part 1 (Up to 12 + 1)

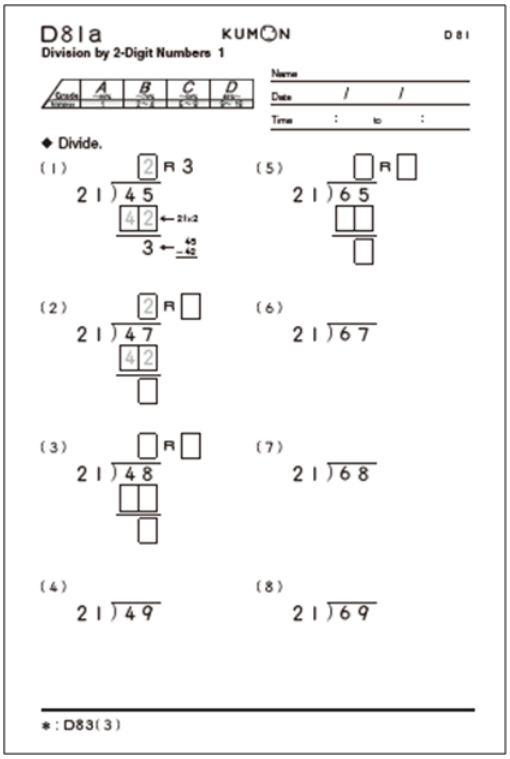
Name
Date / /
Time : to :

◆ Write the number that comes next.

2 ⟶ 3
2 + 1 = 3
Two plus one equals three.

4 ⟶ ☐
4 + 1 = ☐
Four plus one equals

5 ⟶ ☐
5 + 1 = ☐



D81a KUMON D 81
Division by 2-Digit Numbers 1

Name
Date / /
Time : to :

◆ Divide.

(1) 2 R 3
21)45
42 ← 21×2
3 ← 45 − 42

(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)

Appendix B: Data Cleaning and Merging

Sample Attrition: Table B1 shows that the baseline test scores are not correlated with the probability of being out of sample in the endline.

Table B1— Characteristics of dropouts and the sample used in the analysis

Dep. Var	Dropout Grade 3 OLS (1)	Dropout Grade 3 Probit (2)	Dropout Grade 4 OLS (3)	Dropout Grade 4 Probit (4)
Baseline DT Score	0.001	0.009	-0.006	-0.022
	(0.005)	(0.036)	(0.012)	(0.049)
Baseline PTSII-C Score	-0.001	-0.012	-0.002	-0.009
	(0.001)	(0.009)	(0.002)	(0.013)
Number of Observations	481	481	357	357
R-squared	0.008		0.017	

Notes: Asymptotic standard errors are shown in parentheses and are clustered at the school level (34 clusters). The asterisks reflect the significance levels obtained by a clustered wild bootstrap-t procedure; ***, **, and * denote the 1 percent, 5 percent, and 10 percent levels, respectively.

Data Merging: We use student number and school number, which are uniquely assigned to each student and each school in our experiment, to merge the different datasets.

Table B2— Summary Statistics

Variables	Baseline				Endline			
	Control Mean	Treatment Mean	Difference	Observations	Control Mean	Treatment Mean	Difference	Observations
DT Score	45.96	46.00	-0.05	968	56.24	63.97	-7.73***	811
	(17.36)	(17.13)	(0.55)		(14.33)	(10.43)	(0.87)	
DT Time	9.97	9.53	0.43***	968	9.05	6.49	2.56***	811
	(0.27)	(1.37)	(0.07)		(1.55)	(2.29)	(0.14)	
DT Score per Minute[a]	4.62	5.05	-0.43***	968	6.53	11.56	-5.03***	811
	(1.77)	(2.45)	(0.14)		(2.73)	(5.75)	(0.32)	
PTSII-C Score[b]	38.77	34.57	4.20***	1,004	47.44	58.15	-10.71***	837
	(15.25)	(10.31)	(0.86)		(12.66)	(13.95)	(0.92)	
RSES [c]	0.00	0.00	0.00	1,004	-0.03	0.01	-0.04	832
	(0.45)	(0.39)	(0.03)		(0.45)	(0.43)	(0.03)	
CPCS [c]	-0.04	0.03	-0.07	1,004	-0.04	0.03	-0.07***	
	(0.43)	(0.37)	(0.03)		(0.43)	(0.40)	(0.03)	
Cheating	0.00	0.05	-0.05***	1,004				
	(0.00)	(0.22)	(0.01)					
Mismanagement of time (Shorter)	0.13	0.06	0.06***	1,004				
	(0.33)	(0.24)	(0.02)					
Mismanagement of time (Unlimited)	0.13	0.06	0.07***	1,004				
	(0.33)	(0.24)	(0.02)					
Longer Session					-	0.34	-	526
						(0.48)		
Demographic								
Female	0.62	0.58	0.04	974				
	(0.49)	(0.49)	(0.03)					

Notes: Standard deviations for the columns on the means of control and treatment groups and asymptotic standard errors for the column on Difference are shown in parentheses and are clustered at the school level (34 clusters). The asterisks reflect the significance levels obtained by a clustered wild bootstrap-t procedure; ***, **, and * denote the 1 percent, 5 percent, and 10 percent levels, respectively.

[a]: DT Score per Minute stands for the Diagnostic (Math) Test Score per minute: 70 questions must be solved correctly in 10 minutes.

[b]: PTSII-C Score stands for the Proficiency Test of Self Learning first half score, which consists of 348 math questions.

[c]: Among the second half of the Proficiency Test of Self Learning, consisting of 27 survey questions prepared by Kumon, 10 are consistent with the Children's Perceived Competence Scale (CPCS Non-cognitive Score) and 8 with the Rosenberg Self-Esteem Scale (RSES Non-cognitive Score). For survey questions related to each Non-cognitive Score, see Appendix C. The responses are recorded on a four-point scale: 1=Strongly Agree, 2=Somewhat Agree, 3=Somewhat Disagree, 4=Strongly Disagree. Both cognitive and non-cognitive test scores are standardized and used in the regression analysis.]

Appendix C: Non-cognitive Ability Survey Questions

Table C1— PTS II survey questions for measuring non-cognitive abilities

Number	Question in English	CPCS	RSES	GRIT
1	I did well on this test.			
2	I can do most things better than other people.	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 about 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.			x
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	When studying, I believe everything will go well if I correctly follow 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 around me 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 around me who I don't want to lose against. b. There is someone around me who I am always competing with.			
26	I always try to do something when things don't go as expected.			x
27	It doesn't matter whether I fail in the beginning because I believe that things will eventually work out.			x

Note: Among the 27 survey questions prepared by Kumon, 10 are consistent with the Children's Perceived Competence Scale; CPCS (Sakurai and Matsui (1992) Harter (1979)), 8 with the Rosenberg Self-Esteem Scale; RSES (Rosenberg (1965)), and 3 with the Grit Scale; GRIT (Duckworth et al. (2007)). The rest are more specific to the Kumon method of learning original with four Bangladesh-specific questions (questions 24-27). The Japanese version of the original Kumon survey questions is based on Sakurai and Matsui (1992).

Another measurement we consider is the variance in the difference between the standardized value of teacher evaluations and students' actual math test scores. A reduction in this variance implies that the teacher is able to more accurately

Table C2— Dependent Variable: Variance of Difference between Teacher Assessment and Actual Test Scores

Dependent Variables	All Sample	Grade 3 Students	Grade 4 Students
DT Score			
Treatment	0.080	0.222	-0.062
	(0.089)	(0.137)	(0.098)
Endline	-0.057	-0.177*	0.144
	(0.086)	(0.086)	(0.144)
Treatment*Endline	-0.074	-0.073	-0.170
	(0.128)	(0.158)	(0.203)
Constant	1.127***	1.125***	1.129***
	(0.062)	(0.074)	(0.106)
Number of Observations	64	36	28
R-squared	0.047	0.357	0.094
PTSII-C Score			
Treatment	0.069	0.145	-0.018
	(0.072)	(0.104)	(0.100)
Endline	-0.005	-0.004	0.004
	(0.076)	(0.072)	(0.169)
Treatment*Endline	-0.153	-0.178	-0.152
	(0.098)	(0.112)	(0.193)
Constant	1.069***	1.052***	1.089***
	(0.052)	(0.061)	(0.092)
Number of Observations	66	38	28
R-squared	0.064	0.146	0.064

Notes: Asymptotic standard errors are shown in parentheses and are clustered at the school level (34 clusters). The asterisks reflect significance levels obtained by a clustered wild bootstrap-t procedure; ***, **, and * denote the 1 percent, 5 percent, and 10 percent levels, respectively.

track students' math ability, as measured by the DT score per minute and PTSII-C score, thus signifying an improvement in their assessment ability over time. For this measurement, we first standardize both the teacher evaluations and actual math test scores and calculate the school-level variance in the difference between these two values. We then employ the difference-in-differences framework.

Table C2 reports the changes in teacher assessment as shown by the precision measure, taking the variance between the difference in standardized teachers' evaluation and standardized student and student cognitive-test scores. In the "treatment" for coefficient of interest, the interaction term between the treatment and the time dummy in the difference-in-differences is specified, so the signs are consistent across all grades and for both DT score per minute and PTSII-C score, while no grades demonstrate significant results. Overall, the findings suggest that

teacher assessment ability of students' math skills show some improvement, but the significance level varies by grade and type of test.

Appendix D: ANCOVA Results

As a robustness check, we use an ANCOVA model, which allows us to estimate the causal effect of a program by comparing outcomes in the treatment group with outcomes in the control group while controlling for the value of the outcome variable (and other relevant predictors) at the baseline. Hence, we minimize any potential sampling error in the impact estimates.

Unlike the case of a canonical difference-in-differences analysis, ANCOVA analyses are less sensitive to natural within-person variation in the baseline and endline variables McKenzie (2012).

Table D1— Impact of Kumon on Students' Cognitive Abilities: ANCOVA Estimates

Dependent Variables	DT Score per min[a] (1)	DT Score (2)	DT Time (3)	PTSII-C Score[b] (4)	RSES (5)	CPCS (6)
Treatment	2.172***	0.478***	-2.312***	1.037***	-0.010	0.076
	(0.514)	(0.126)	(0.506)	(0.191)	(0.141)	(0.135)
Constant	1.134	0.631***	-0.961	0.558*	-0.022	-0.036
	(0.296)	(0.123)	(0.337)	(0.124)	(0.103)	(0.098)
Num of Obs.	799	799	799	837	832	832
R-squared	0.254	0.162	0.322	0.321	0.031	0.040

Notes: Asymptotic standard errors are shown in parentheses and are clustered at the school level (34 clusters). The asterisks reflect significance levels obtained by a clustered wild bootstrap-t procedure; ***, **, and * denote the 1 percent, 5 percent, and 10 percent levels, respectively.

[a]: DT Score per Minute stands for Diagnostic (Math) Test Score per minute: 70 questions must be solved correctly in 10 minutes. Both cognitive and non-cognitive test scores are standardized and used in the regression analysis.

[b]: PTSII-C Score stands for the Proficiency Test of Self Learning first half score, which consists of 348 math questions.

Table D2— Heterogenous Impact of Kumon on Students' Cognitive Abilities: ANCOVA Estimates

	Initial RSES Score					Initial CPCS Score				
Dependent Variables	DT Score per min[a]	DT Score	DT Time	PTSII-C Score[b]	RSES	DT Score per min[a]	DT Score	DT Time	PTSII-C Score[b]	CPCS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: High Initial Cognitive High Initial Non-cognitive										
Treatment	3.248***	0.387**	-3.721***	1.206***	-0.074	3.291***	0.386***	-3.539***	1.131***	0.023
	(0.728)	(0.142)	(0.584)	(0.262)	(0.174)	(0.842)	(0.130)	(0.646)	(0.277)	(0.178)
Constant	0.874**	0.521***	-0.801**	0.528***	0.176	0.922*	0.643***	-1.209***	0.504**	0.141
	(0.352)	(0.121)	(0.313)	(0.173)	(0.160)	(0.479)	(0.155)	(0.435)	(0.201)	(0.155)
Num of Obs.	186	186	186	189	189	188	188	188	188	188
R-squared	0.368	0.119	0.478	0.347	0.027	0.373	0.114	0.450	0.354	0.046
Panel B: High Initial Cognitive Low Initial Non-cognitive										
Treatment	3.092***	0.270	-3.730***	1.083***	-0.124	2.901***	0.238	-3.815***	1.249***	-0.125
	(0.897)	(0.161)	(0.715)	(0.283)	(0.216)	(0.790)	(0.217)	(0.680)	(0.248)	(0.216)
Constant	0.982**	0.772***	-1.101**	0.411**	0.102	1.118***	0.645***	-0.630**	0.414**	0.046
	(0.414)	(0.168)	(0.465)	(0.190)	(0.205)	(0.325)	(0.174)	(0.298)	(0.158)	(0.179)
Num of Obs.	169	169	169	180	180	167	167	167	181	181
R-squared	0.415	0.107	0.531	0.323	0.116	0.370	0.085	0.540	0.312	0.047
Panel C: Low Initial Cognitive High Initial Non-cognitive										
Treatment	2.450**	0.579**	-2.376***	1.484***	0.032	2.248**	0.465	-2.195***	1.320***	0.241
	(0.956)	(0.277)	(0.695)	(0.277)	(0.264)	(1.011)	(0.283)	(0.755)	(0.302)	(0.314)
Constant	1.402**	0.539*	-0.939**	0.256	-0.112	1.533**	0.571**	-0.878**	0.422	-0.057
	(0.622)	(0.291)	(0.369)	(0.210)	(0.174)	(0.725)	(0.266)	(0.387)	(0.272)	(0.262)
Num of Obs.	150	150	150	150	150	141	141	141	150	150
R-squared	0.188	0.147	0.310	0.372	0.032	0.151	0.104	0.258	0.330	0.041
Panel D: Low Initial Cognitive Low Initial Non-cognitive										
Treatment	1.196***	0.546***	-1.532***	1.138***	0.372**	1.467***	0.650***	-1.760***	1.248***	0.252
	(0.347)	(0.181)	(0.501)	(0.226)	(0.169)	(0.309)	(0.172)	(0.484)	(0.254)	(0.190)
Constant	1.135***	0.655***	-0.841**	0.778***	-0.160	1.078***	0.627***	-0.858*	0.627***	-0.344**
	(0.326)	(0.175)	(0.372)	(0.183)	(0.152)	(0.305)	(0.172)	(0.424)	(0.155)	(0.128)
Num of Obs.	171	171	171	177	177	180	180	180	177	177
R-squared	0.159	0.142	0.275	0.376	0.060	0.212	0.189	0.315	0.400	0.090

Notes: Asymptotic standard errors are shown in parentheses and are clustered at the school level (34 clusters). The asterisks reflect the significance levels obtained by a clustered wild bootstrap-t procedure; ***, **, and * denote the 1 percent, 5 percent, and 10 percent levels, respectively.

[a]: DT Score per Minute stands for Diagnostic (Math) Test Score per minute: 70 questions must be solved correctly in 10 minutes. Both cognitive and non-cognitive test scores are standardized and used in the regression analysis.

[b]: PTSII-C Score stands for the Proficiency Test of Self Learning first half score, which consists of 348 math questions.

Table D3— Impact of Kumon on Students' Cognitive and Non-cognitive Abilities: Estimates Controlling for Longer Sessions - ANCOVA Estimates

Dependent Variables	DT Score per min[a] (1)	DT Score (2)	DT Time (3)	PTSII-C Score[b] (4)	RSES [c] (5)	CPCS [c] (6)
Treatment	2.424***	0.396***	-2.719***	1.046***	0.008	0.122
	(0.648)	(0.135)	(0.548)	(0.226)	(0.171)	(0.166)
Treatment x Longer session	-0.761	0.246	1.227**	-0.029	-0.053	-0.136
	(0.532)	(0.129)	(0.537)	(0.230)	(0.205)	(0.200)
Constant	1.128***	0.633***	-0.949***	0.558***	-0.023	-0.039
	(0.287)	(0.120)	(0.318)	(0.124)	(0.102)	(0.096)
Num of Obs.	799	799	799	837	832	832
R-squared	0.264	0.173	0.352	0.321	0.031	0.041

Notes: Asymptotic standard errors are shown in parentheses and are clustered at the school level (34 clusters). The asterisks reflect the significance levels obtained by a clustered wild bootstrap-t procedure; ***, **, and * denote the 1 percent, 5 percent, and 10 percent levels, respectively.
[a]: DT Score per Minute stands for the Diagnostic (Math) Test Score per minute: 70 questions must be solved correctly in 10 minutes.
[b]: PTSII-C Score stands for the Proficiency Test of Self Learning first half score, which consists of 348 math questions.
[c]: Among the second half of the Proficiency Test of Self Learning, consisting of 27 survey questions prepared by Kumon, 10 are consistent with the Children's Perceived Competence Scale (CPCS Non-cognitive Score) and 8 with the Rosenberg Self-Esteem Scale (RSES Non-cognitive Score). For survey questions related to each Non-cognitive Score, see Appendix C. The responses are recorded on a four-point scale: 1=Strongly Agree, 2=Somewhat Agree, 3=Somewhat Disagree, 4=Strongly Disagree. Both cognitive and non-cognitive test scores are standardized and used in the regression analysis.]

Appendix E: Heterogenous Treatment with Continuous Cognitive/Non-cognitive Scores

Table E1— Impact of Kumon on Students' Cognitive and Non-cognitive Abilities (continuous measures): DID Estimates

Dependent Variables	Initial RSES Score					Initial CPCS Score				
	DT Score per min [a] (1)	DT Score (2)	DT Time (3)	PTSII-C Score [b] (4)	RSES [c] (5)	DT Score per min [a] (6)	DT Score (7)	DT Time (8)	PTSII-C Score [b] (9)	CPCS [c] (10)
Treatment	2.155***	0.577***	-2.158***	1.186***	0.007	2.133***	0.580***	-2.142***	1.182***	-0.037
	(0.506)	(0.163)	(0.509)	(0.197)	(0.174)	(0.496)	(0.163)	(0.500)	(0.196)	(0.180)
Treatment x Initial Cognitive Score	-0.496***	-0.920***	-0.910***	-0.376***	0.059	-0.517***	-0.915***	-0.919***	-0.383***	0.033
(continuous) [d]	(0.101)	(0.054)	(0.082)	(0.102)	(0.107)	(0.094)	(0.055)	(0.081)	(0.104)	(0.104)
Treatment x Initial Non-cognitive Score	0.261	-0.036	-0.151	-0.082	-0.933***	0.314	-0.050	-0.253	-0.047	-0.895***
(continuous) [e]	(0.276)	(0.044)	(0.236)	(0.068)	(0.080)	(0.288)	(0.038)	(0.246)	(0.068)	(0.079)
Treatment x Initial Cognitive Score	0.067	-0.048	-0.042	-0.072	0.045	0.156	-0.039	-0.010	-0.008	0.107*
x Initial Non-cognitive Score	(0.155)	(0.040)	(0.138)	(0.076)	(0.069)	(0.137)	(0.032)	(0.138)	(0.069)	(0.057)
Constant	1.088***	0.503***	-1.007***	0.519***	0.000	1.093***	0.502***	-1.016***	0.518***	0.066
	(0.318)	(0.155)	(0.344)	(0.130)	(0.082)	(0.321)	(0.155)	(0.345)	(0.130)	(0.087)
Num of Obs.	799	799	799	787	696	799	799	799	787	696
R-squared	0.239	0.415	0.331	0.450	0.246	0.242	0.415	0.334	0.449	0.228

Notes: Asymptotic standard errors are shown in parentheses and are clustered at the school level (34 clusters). The asterisks reflect the significance levels obtained by a clustered wild bootstrap-t procedure; ***, **, and * denote the 1 percent, 5 percent, and 10 percent levels, respectively.

[a]: DT Score per Minute stands for Diagnostic (Math) Test Score per minute: 70 questions must be solved correctly in 10 minutes. Both cognitive and non-cognitive test scores are standardized and used in the regression analysis.

[b]: PTSII-C Score stands for the Proficiency Test of Self Learning first half score, which consists of 348 math questions.

[c]: Among the second half of the Proficiency Test of Self Learning, consisting of 27 survey questions prepared by Kumon, 10 are consistent with the Children's Perceived Competence Scale (CPCS Non-cognitive Score) and 8 with the Rosenberg Self-Esteem Scale (RSES Non-cognitive Score). For survey questions related to each Non-cognitive Score, see Appendix C. The responses are recorded on a four-point scale: 1=Strongly Agree, 2=Somewhat Agree, 3=Somewhat Disagree, 4=Strongly Disagree. Both cognitive and non-cognitive test scores are standardized and used in the regression analysis.]

[d]: The Initial Cognitive Score stands for the DT Score for columns (1)-(3), (6)-(8) and the PTSII-C Score for columns (4),(5),(9),(10).

[e]: The Initial Non-cognitive Score stands for RSES for columns (1)-(5). For columns (6)-(10), CPCS is used.

Appendix F: Endline Tables

Table F1— Impact of Kumon on Students' Cognitive Abilities: Endline Estimates

Dependent Variables	DT Score per min[a] (1)	DT Score (2)	DT Time (3)	PTSII-C Score[b] (4)	RSES (5)	CPCS (6)
Treatment	2.146***	0.463***	-2.300***	0.947***	0.003	0.100
	(0.511)	(0.125)	(0.504)	(0.214)	(0.150)	(0.144)
Constant	1.168***	0.655***	-0.976***	0.548***	-0.029	-0.052
	(0.293)	(0.129)	(0.334)	(0.140)	(0.107)	(0.098)
Num of Obs.	811	811	811	837	832	832
R-squared	0.234	0.129	0.321	0.220	0.021	0.031

Notes: Asymptotic standard errors are shown in parentheses and are clustered at the school level (34 clusters). The asterisks reflect significance levels obtained by a clustered wild bootstrap-t procedure; ***, **, and * denote the 1 percent, 5 percent, and 10 percent levels, respectively.

[a]: DT Score per Minute stands for Diagnostic (Math) Test Score per minute: 70 questions must be solved correctly in 10 minutes. Both cognitive and non-cognitive test scores are standardized and used in the regression analysis.

[b]: PTSII-C Score stands for the Proficiency Test of Self Learning first half score, which consists of 348 math questions.

Table F2— Heterogenous Impact of Kumon on Students' Cognitive Abilities: Endline Estimates

	Initial RSES Score					Initial CPCS Score				
Dependent Variables	DT Score per min[a]	DT Score	DT Time	PTSII-C Score[b]	RSES	DT Score per min[a]	DT Score	DT Time	PTSII-C Score[b]	CPCS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: High Initial Cognitive High Initial Non-cognitive										
Treatment	3.262***	0.366**	-3.698***	1.142***	-0.067	3.302***	0.374***	-3.555***	0.988***	0.026
	(0.708)	(0.138)	(0.607)	(0.303)	(0.174)	(0.801)	(0.125)	(0.660)	(0.326)	(0.177)
Constant	1.123***	0.724***	-0.795**	0.947***	0.122	1.439***	0.793***	-1.218***	0.969***	0.101
	(0.306)	(0.120)	(0.316)	(0.222)	(0.130)	(0.384)	(0.131)	(0.438)	(0.214)	(0.129)
Num of Obs.	186	186	186	189	189	188	188	188	188	188
R-squared	0.363	0.102	0.478	0.262	0.025	0.353	0.103	0.450	0.267	0.045
Panel B: High Initial Cognitive Low Initial Non-cognitive										
Treatment	3.088***	0.266*	-3.735***	0.902**	-0.085	2.909***	0.228	-3.810***	1.192***	-0.101
	(0.884)	(0.156)	(0.726)	(0.337)	(0.234)	(0.786)	(0.222)	(0.684)	(0.297)	(0.210)
Constant	1.540***	0.921***	-1.117**	0.766***	-0.179	1.206***	0.848***	-0.631**	0.674***	-0.135
	(0.354)	(0.104)	(0.467)	(0.186)	(0.163)	(0.247)	(0.091)	(0.297)	(0.187)	(0.144)
Num of Obs.	169	169	169	180	180	167	167	167	181	181
R-squared	0.386	0.094	0.522	0.258	0.063	0.369	0.066	0.540	0.259	0.026
Panel C: Low Initial Cognitive High Initial Non-cognitive										
Treatment	2.405**	0.563*	-2.407***	1.448***	0.035	2.183**	0.451	-2.227***	1.294***	0.245
	(0.935)	(0.292)	(0.726)	(0.266)	(0.279)	(0.980)	(0.292)	(0.785)	(0.289)	(0.325)
Constant	1.015**	0.376	-0.953**	-0.009	0.031	1.086**	0.463	-0.892**	0.117	0.033
	(0.461)	(0.306)	(0.380)	(0.184)	(0.142)	(0.474)	(0.276)	(0.399)	(0.211)	(0.224)
Num of Obs.	150	150	150	150	150	141	141	141	150	150
R-squared	0.173	0.107	0.302	0.351	0.015	0.133	0.089	0.250	0.305	0.036
Panel D: Low Initial Cognitive Low Initial Non-cognitive										
Treatment	1.202***	0.554***	-1.532***	1.151***	0.371**	1.475***	0.661***	-1.760***	1.248***	0.249
	(0.349)	(0.176)	(0.499)	(0.238)	(0.167)	(0.311)	(0.169)	(0.483)	(0.262)	(0.193)
Constant	1.020***	0.573***	-0.843**	0.227	-0.182*	0.928***	0.502***	-0.860*	0.155	-0.209**
	(0.294)	(0.174)	(0.371)	(0.174)	(0.093)	(0.277)	(0.172)	(0.423)	(0.189)	(0.100)
Num of Obs.	171	171	171	177	177	180	180	180	177	177
R-squared	0.157	0.132	0.269	0.309	0.060	0.208	0.164	0.308	0.348	0.077

Notes: Asymptotic standard errors are shown in parentheses and are clustered at the school level (34 clusters). The asterisks reflect the significance levels obtained by a clustered wild bootstrap-t procedure; ***, **, and * denote the 1 percent, 5 percent, and 10 percent levels, respectively.

[a]: DT Score per Minute stands for Diagnostic (Math) Test Score per minute: 70 questions must be solved correctly in 10 minutes. Both cognitive and non-cognitive test scores are standardized and used in the regression analysis.

[b]: PTSII-C Score stands for the Proficiency Test of Self Learning first half score, which consists of 348 math questions.

Table F3— Impact of Kumon on Students' Cognitive and Non-cognitive Abilities: Estimates Controlling for Longer Sessions - Endline Estimates

Dependent Variables	DT Score per min[a] (1)	DT Score (2)	DT Time (3)	PTSII-C Score[b] (4)	RSES [c] (5)	CPCS [c] (6)
Treatment	2.409***	0.392***	-2.691***	0.981***	0.030	0.156
	(0.634)	(0.133)	(0.542)	(0.258)	(0.182)	(0.175)
Treatment x Longer session	-0.800	0.214*	1.193**	-0.104	-0.083	-0.169
	(0.531)	(0.126)	(0.526)	(0.266)	(0.208)	(0.202)
Constant	1.161***	0.657***	-0.965***	0.547***	-0.030	-0.054
	(0.284)	(0.127)	(0.317)	(0.141)	(0.106)	(0.096)
Num of Obs.	811	811	811	837	832	832
R-squared	0.245	0.138	0.349	0.221	0.022	0.034

Notes: Asymptotic standard errors are shown in parentheses and are clustered at the school level (34 clusters). The asterisks reflect the significance levels obtained by a clustered wild bootstrap-t procedure; ***, **, and * denote the 1 percent, 5 percent, and 10 percent levels, respectively.
[a]: DT Score per Minute stands for the Diagnostic (Math) Test Score per minute: 70 questions must be solved correctly in 10 minutes.
[b]: PTSII-C Score stands for the Proficiency Test of Self Learning first half score, which consists of 348 math questions.
[c]: Among the second half of the Proficiency Test of Self Learning, consisting of 27 survey questions prepared by Kumon, 10 are consistent with the Children's Perceived Competence Scale (CPCS Non-cognitive Score) and 8 with the Rosenberg Self-Esteem Scale (RSES Non-cognitive Score). For survey questions related to each Non-cognitive Score, see Appendix C. The responses are recorded on a four-point scale: 1=Strongly Agree, 2=Somewhat Agree, 3=Somewhat Disagree, 4=Strongly Disagree. Both cognitive and non-cognitive test scores are standardized and used in the regression analysis.]

Appendix G: Results excluding students with DT mismatch at baseline

Table G1— Impact of Kumon on Students' Cognitive Abilities: DID Estimates

Dependent Variables	DT Score per min[a] (1)	DT Score (2)	DT Time (3)	PTSII-C Score[b] (4)	RSES (5)	CPCS (6)
Treatment	2.174***	0.505*	-2.269***	1.198***	-0.043	-0.113
	(0.534)	(0.208)	(0.515)	(0.210)	(0.178)	(0.180)
Constant	1.109***	0.583***	-0.955***	0.565***	0.017	0.102
	(0.321)	(0.167)	(0.340)	(0.140)	(0.092)	(0.101)
Num of Obs.	663	663	663	659	570	570
R-squared	0.197	0.063	0.218	0.435	0.022	0.021

Notes: Asymptotic standard errors are shown in parentheses and are clustered at the school level (34 clusters). The asterisks reflect significance levels obtained by a clustered wild bootstrap-t procedure; ***, **, and * denote the 1 percent, 5 percent, and 10 percent levels, respectively.

[a]: DT Score per Minute stands for Diagnostic (Math) Test Score per minute: 70 questions must be solved correctly in 10 minutes. Both cognitive and non-cognitive test scores are standardized and used in the regression analysis.

[b]: PTSII-C Score stands for the Proficiency Test of Self Learning first half score, which consists of 348 math questions.

Table G2— Heterogenous Impact of Kumon on Students' Cognitive Abilities: DID Estimates Excluding Students with DT Mismatch at Baseline

	Initial RSES Score					Initial CPCS Score				
Dependent Variables	DT Score per min[a]	DT Score	DT Time	PTSII-C Score[b]	RSES	DT Score per min[a]	DT Score	DT Time	PTSII-C Score[b]	CPCS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: High Initial Cognitive High Initial Non-cognitive										
Treatment	3.219***	0.453**	-3.396***	1.254***	0.044	3.298***	0.453**	-3.341***	1.240***	0.075
	(0.767)	(0.183)	(0.639)	(0.254)	(0.203)	(0.876)	(0.173)	(0.692)	(0.267)	(0.196)
Constant	0.332	-0.116	-0.721**	0.217	-0.693***	0.634	-0.028	-1.089**	0.148	-0.704***
	(0.313)	(0.125)	(0.307)	(0.232)	(0.154)	(0.391)	(0.137)	(0.434)	(0.224)	(0.146)
Num of Obs.	157	157	157	164	164	147	147	147	158	158
R-squared	0.342	0.116	0.369	0.483	0.037	0.319	0.119	0.333	0.483	0.028
Panel B: High Initial Cognitive Low Initial Non-cognitive										
Treatment	3.097***	0.286	-3.716***	1.248***	-0.190	2.843***	0.265	-3.603***	1.306***	-0.209
	(0.922)	(0.192)	(0.706)	(0.219)	(0.235)	(0.792)	(0.224)	(0.700)	(0.226)	(0.290)
Constant	0.840**	0.170	-1.065**	0.087	0.588***	0.523*	0.074	-0.663**	0.152	0.663***
	(0.372)	(0.136)	(0.460)	(0.149)	(0.175)	(0.259)	(0.106)	(0.307)	(0.163)	(0.192)
Num of Obs.	124	124	124	146	146	134	134	134	152	152
R-squared	0.345	0.084	0.462	0.448	0.059	0.331	0.068	0.471	0.451	0.034
Panel C: Low Initial Cognitive High Initial Non-cognitive										
Treatment	2.480**	0.628**	-2.504***	1.538***	0.020	2.290**	0.549**	-2.323**	1.354***	0.212
	(0.942)	(0.244)	(0.803)	(0.293)	(0.232)	(0.978)	(0.257)	(0.866)	(0.324)	(0.288)
Constant	1.656***	1.033***	-0.995**	0.648***	-0.674***	1.829***	1.229***	-0.935**	0.818***	-0.656***
	(0.434)	(0.217)	(0.411)	(0.212)	(0.147)	(0.455)	(0.170)	(0.432)	(0.249)	(0.222)
Num of Obs.	120	120	120	119	119	111	111	111	112	112
R-squared	0.203	0.118	0.227	0.387	0.022	0.179	0.098	0.182	0.342	0.065
Panel D: Low Initial Cognitive Low Initial Non-cognitive										
Treatment	1.158***	0.463*	-1.532***	1.134***	0.412**	1.434***	0.589***	-1.760***	1.248***	0.230
	(0.378)	(0.254)	(0.501)	(0.226)	(0.165)	(0.335)	(0.207)	(0.484)	(0.253)	(0.245)
Constant	1.904***	1.503***	-0.846**	0.983***	0.537***	1.704***	1.311***	-0.863*	0.882***	0.589***
	(0.296)	(0.187)	(0.372)	(0.174)	(0.099)	(0.282)	(0.190)	(0.424)	(0.186)	(0.122)
Num of Obs.	143	143	143	141	141	152	152	152	148	148
R-squared	0.101	0.071	0.133	0.342	0.069	0.140	0.120	0.169	0.363	0.024

Notes: Asymptotic standard errors are shown in parentheses and are clustered at the school level (34 clusters). The asterisks reflect the significance levels obtained by a clustered wild bootstrap-t procedure; ***, **, and * denote the 1 percent, 5 percent, and 10 percent levels, respectively.

[a]: DT Score per Minute stands for Diagnostic (Math) Test Score per minute: 70 questions must be solved correctly in 10 minutes. Both cognitive and non-cognitive test scores are standardized and used in the regression analysis.

[b]: PTSII-C Score stands for the Proficiency Test of Self Learning first half score, which consists of 348 math questions.

Table G3— Impact of Kumon on Students' Cognitive and Non-cognitive Abilities: Estimates Controlling for Longer Sessions - DID Estimates

Dependent Variables	DT Score per min[a]	DT Score	DT Time	PTSII-C Score[b]	RSES [c]	CPCS [c]
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	2.397***	0.299	-2.790***	1.229***	-0.183	-0.237
	(0.769)	(0.243)	(0.620)	(0.266)	(0.190)	(0.199)
Treatment x Longer session	-0.669	0.618**	1.562*	-0.093	0.535	0.474
	(0.799)	(0.295)	(0.773)	(0.301)	(0.320)	(0.315)
Constant	1.103***	0.589***	-0.941***	0.563***	0.011	0.096
	(0.313)	(0.162)	(0.320)	(0.141)	(0.089)	(0.098)
Num of Obs.	663	663	663	659	570	570
R-squared	0.203	0.090	0.254	0.436	0.034	0.030

Notes: Asymptotic standard errors are shown in parentheses and are clustered at the school level (34 clusters). The asterisks reflect the significance levels obtained by a clustered wild bootstrap-t procedure; ***, **, and * denote the 1 percent, 5 percent, and 10 percent levels, respectively.

[a]: DT Score per Minute stands for the Diagnostic (Math) Test Score per minute: 70 questions must be solved correctly in 10 minutes.

[b]: PTSII-C Score stands for the Proficiency Test of Self Learning first half score, which consists of 348 math questions.

[c]: Among the second half of the Proficiency Test of Self Learning, consisting of 27 survey questions prepared by Kumon, 10 are consistent with the Children's Perceived Competence Scale (CPCS Non-cognitive Score) and 8 with the Rosenberg Self-Esteem Scale (RSES Non-cognitive Score). For survey questions related to each Non-cognitive Score, see Appendix C. The responses are recorded on a four-point scale: 1=Strongly Agree, 2=Somewhat Agree, 3=Somewhat Disagree, 4=Strongly Disagree. Both cognitive and non-cognitive test scores are standardized and used in the regression analysis.]

Appendix H: Teacher Assessment Ability

Table H1— Association between Teachers' Assessment and Student Performance

Dependent Variables	DT Score per minute[a]	PTSII-C Score[b]
Teacher_evaluation* (1-Treatment)*(1-Endline)	0.386***	0.124
	(0.055)	(0.201)
Teacher_evaluation*Treatment*(1-Endline)	0.479***	0.306***
	(0.064)	(0.049)
Teacher_evaluation*(1-Treatment)*Endline	0.177**	0.155
	(0.085)	(0.122)
Teacher_evaluation*Treatment*Endline	0.623***	0.545***
	(0.222)	(0.083)
Control_Baseline = Treatment_Baseline	1.22	0.77
Control_Endline = Treatment_Endline	3.53*	0.01**
Num of Obs.	1,268	1,292
R-squared	0.531	0.532

Notes: Asymptotic standard errors are shown in parentheses and are clustered at the school level (34 clusters). The asterisks reflect the significance levels obtained by a clustered wild bootstrap-t procedure; ***, **, and * denote the 1 percent, 5 percent, and 10 percent levels, respectively.
[a]: DT Score per Minute stands for the Diagnostic (Math) Test Score per minute: 70 questions must be solved correctly in 10 minutes.
[b]: PTSII-C Score stands for the Proficiency Test of Self Learning first half score, which consists of 348 math questions.