CIRJE-F-1127

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

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
University College London

October 2019; Revised in November 2020

Fighting the 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*

November, 2020

Abstract

This paper investigates the effectiveness of a globally popular method of self-learning at the right level in improving learning outcomes—the cognitive and non-cognitive abilities of disadvantaged pupils—in a developing country, Bangladesh. Using a randomized control trial design, we find substantial improvements in cognitive abilities measured by mathematics test scores and catch-up effects on aspects of non-cognitive abilities or personality traits measured by a self-esteem scale. We also find a longer-term impact on cognitive abilities regarding the math scores students obtained on national-level exams compared to the baseline test scores. Moreover, teachers' abilities to assess students' performance substantially improve. Our estimates indicate that the program's benefits exceed its costs. The above findings suggest that self-learning at the right level can effectively supplement the quality of primary education and hence address the learning crisis in developing countries.

JEL code: 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, 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 College London, Drayton House, 30 Gordon Street, Kings Cross, London WC1H 0AX, United Kingdom (e-mail: hikaru.kawarazaki.20@ucl.ac.uk). The opinions expressed in this article are the authors' 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.

1 Introduction

The learning crisis refers to the global phenomenon where over 60 percent of children who complete their primary education in low- and middle-income countries fail to achieve a minimum proficiency in mathematics and reading (World Bank, 2018; UNESCO, 2013). Given that education is an important link to all sustainable development goals (SDGs), improving the quality of education is a *sine qua non* for achieving SDGs (United Nations, 2018). Teaching at the right level (TaRL) programs are gaining increasing 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).[1] For example, Muralidharan, Singh and Ganimian (2019) in India found that individualized technology-aided instruction programs can improve test scores. The lack of appropriate infrastructure in developing countries, however, can potentially constrain the use of such effective programs.

We evaluate the effectiveness of an individualized self-learning program, the Kumon method of learning (hereafter, Kumon), which does not necessarily rely on the use of Information and Communication Technology (ICT), in supplementing the learning quality of primary schools in Bangladesh. This is a globally popular non-formal education program that is based on a paper-and-pencil approach designed to ensure that each student always studies at the level that is "just right" for them.[2] In Kumon, each student begins at an individually suitable starting point and learns new concepts in small steps in which learning is enforced through easily understandable hints and examples.

Bangladesh has successfully increased school enrollment and narrowed gender gaps. Non-formal education has been critical to this process in addition to conventional public formal education. On the non-formal side, NGOs such as BRAC have played an important role in collaboration with the government. In particular, BRAC primary schools (BPSs) have provided disadvantaged students with a four-year accelerated program that covers the five-year public primary school curriculum.[3] Given the success of BPS in enrollment and reducing primary school dropouts, the government of Bangladesh has scaled up a modified version of BPS under the Reaching Out of School 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 learning remain a serious concern in Bangladesh, as in other developing countries.[4]

[1]Regarding 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).

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

[3]BPS is known as one of the largest and most successful non-formal education programs targeted at 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 provided in Section 2.

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

In this context, we adopt and evaluate the impact of 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 with formal schools (Ahmad and Haque, 2011). BPSs have 30 students per class with diverse backgrounds and a large variance in ability in the subjects taught, particularly mathematics (Nath, 2012). This creates a potential mismatch between the teaching level and students' individual abilities. However, BPSs cannot effectively offer TaRL, as they follow the same instructional approach as public schools. Kumon, as a supplementary approach, could at least partially respond to this mismatch and improve learning outcomes by providing self-learning mathematics materials for each student.[5]

Our findings indicate that Kumon substantially improves students' cognitive abilities as measured through math test scores. 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 tests as measures of cognitive ability. The magnitude of the impact measured by test score per minute is a 2.073 standard deviation, whereby the impact comes through both test score gains and reductions in problem-solving speed.[6] In the case of the time-unadjusted test scores, the magnitude of the impact ranges from 0.501 to 1.212. In terms of non-cognitive abilities measured through certain personality traits, we find catch-up effects among pupils with initially lower abilities compared to the median. We also find a longer-term impact on cognitive ability in terms of math scores obtained in the national-level Primary School Certificate (PSC) examination compared to the baseline test score. Additionally, we find that the intervention significantly improves teachers' abilities to assess their students' performance.

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 presents the econometric evaluation framework, followed by the empirical results. Section 4 compares the benefits and costs of this intervention, and Section 5 concludes the paper.

---

and the likelihood of providing more than 75 percent correct answers was only 9 percent higher when compared with 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 show that the predictive power of non-cognitive abilities is comparable to or exceeds that of cognitive skills in explaining education, success in the labor market, or other outcomes (Heckman, 2006; Heckman, Humphries and Kautz, 2014). Notwithstanding that Kumon has been regarded as a successful non-formal education program in developed countries, it is worth evaluating its impacts on learning outcomes in a disadvantaged setting in a developing country context.

[6]These effects are largely comparable 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. Further, 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.

2 Experiment Design, Data, and Balancing Test

2.1 Setting: BRAC Primary School

Primarily, BPS targets children from disadvantaged social backgrounds who could not access formal schooling at the right age or have dropped out of the 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 that has worked for wages for at least 100 days" and that live within a two-kilometer radius of the school are admitted to BPS (Afroze, 2012). BPS covers the same standard curriculum as public schools. Although the BPS and government primary schools teach the same competency-based curriculum, there are some basic differences between them. Unlike the five-year standard primary school system, BPS offers an accelerated four-year program to help these children readapt to formal education.(Asadullah, 2016). In particular, BPS teachers address students who are falling behind in the following manner: the entry age for students in BPS is higher than that in standard primary schools (the official age is 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, which results in higher contact hours per primary cycle compared with government primary schools; the average class size in BPS (25-30 students) is smaller than 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 lectures followed by assignments. Students are required to pass the grade five terminal examination set by the government (i.e., PSC), which also suggests that BPS provides learners with the same skills that are taught in government schools, whereby teaching to the test potentially affects students' learning.

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

2.2 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 appropriate for them. Students are assigned to an initial level based on their individual performance in a diagnostic test (DT) provided by the Kumon Institute of Education Co., Ltd. rather than on the basis of their school grade or age. The Kumon method is uniquely designed so that the initial level is slightly lower than the student's concurrent maximum capacity in order to: i) ensure that students fully understand the basic concepts and develop a firm foundation for the development of their cognitive abilities and ii) motivate students to continue studying, which also aids the development of their non-cognitive abilities, such as self-esteem and sense of competence. Kumon worksheets are designed, ranging from simple counting to advanced mathematics, with the level of difficulty increasing gradually. Worksheets contain example questions with hints that help students acquire step-by-step problem-solving skills independently.[7] Kumon instructors do not conduct lectures; they simply observe students' progress. They adjust the level of the worksheets if students are stuck on the same worksheet or are unable to 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 materials at an early age. Importantly, slower learners can spend more time on basics without being rushed on to advanced-level materials beyond their level of understanding.

Another feature of Kumon is a system that tracks each student's progress and achievements using personalized record books. Kumon instructors do not teach in class, and hence, do not need extensive prior experience in conducting daily quizzes to monitor each student's understanding and progress. This is because Kumon worksheets are presented in small steps that enable students to learn independently, and there is a set standard time to solve each worksheet, which allows teachers to determine the level students are permitted to advance to or whether they should repeat a level. Detailed progress reports on the worksheets allows instructors to obtain more objective information about their students' abilities and understanding of the mathematics involved.

2.3 Experimental Design

To identify the causal effects of Kumon on young students' learning and their cognitive abilities in particular, we implemented a Randomized Control Trial (RCT) design. Consistent with the effect size of education intervention elsewhere,[8] we hypothesize a minimum detectable effect of 0.40 standard deviation on students' cognitive ability. 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. This gives us a sample of approximately 26 clusters with a statistical power of 0.80. To ensure that we did not lose statistical power due to attrition or other factors, we selected a cluster size of 34 to increase the total student sample, with an average of 30 students per cluster, which gives us a final sample of approximately 1,000 students. We then randomly selected 34 schools from a list of 179 eligible BPSs (located in Dhaka and surrounding areas) for our study, dividing them equally as 17 treatment and 17 control schools. The resulting sample breakdown by class/grade is as follows: 19 (out of 48 schools) for the third grade and 15 (out of 131 schools) for the fourth grade. The schools do not overlap in terms of grade. In other words, in a particular school, we only offer the intervention to either grade three or grade four.

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

[7] See example worksheets of Kumon in Appendix A.

[8] Considering the results from some studies of high-impact education interventions involving teaching at the right level, such as Lakshminarayana et al. (2013) and Duflo, Dupas and Kremer (2011)

Unlike the regular Kumon sessions elsewhere, we did not require students to complete related homework to restrict the daily 30-minutes regular Kumon learning sessions. In addition, unlike a standard Kumon center that usually offers sessions outside school, our treatment school students remained in the classroom in which their regular BPS classes were held. BPSs run for six days a week, except on public holidays, teacher refreshment days, 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. provided an intervention package consisting of a mathematics materials set and an instructor manual with sheets for the BRAC teacher.[9] The full material set consisted of i) mathematics worksheets with questions at various difficulty levels and achievement tests at the end of each level; and ii) a grading notebook to record students' daily progress, including the level of worksheet that a student worked on, the number of repetitions required before achieving a full score on the worksheet, and the number of worksheets that students finally completed.[10]

During the administration of the Kumon program, the BPS teachers did not conduct lectures; they simply observed their students' progress. They only intervened when students were stuck on the same worksheet or could not find the right answer after many attempts. They adjust the level of worksheets in such cases. The BPS teachers also provided guidance when advanced students proceeded to entirely new materials beyond the regular curriculum. The marking assistants helped the teachers grade and record the worksheets. Until the session ended, students either moved on to a new worksheet once they had achieved a full score on the previous one or continued to try and correct their answers until they achieved a full score within the designated time frame.

2.4 Data Description

We constructed cognitive ability measures at both the baseline and endline based on two different mathematics test scores for both the treatment and control school students. These mathematics tests are the Diagnostic Test (DT) and Proficiency Test of Self Learning (PTSII). 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 is time-specific and requires students to answer 70 questions within a maximum of 10 minutes. Hence, for the DT, we show test scores per minute (DT Score per min) to determine students' cognitive abilities. 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 measuring aspects of non-cognitive abilities (see Table B1 of Appendix B). Among the 27 questions, 8 are consistent with the Rosenberg Self-Esteem Scale (RSES

[9]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 were 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 were held during the implementation period. Two marking assistants (graders) were provided for each class to support the grading and recording of worksheets during the Kumon session. BPS teachers monitored students and determined the level of worksheets that students were required to work on.

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

Index) (Rosenberg, 1965), and 10 are consistent with the Children's Perceived Competence Scale (CPCS Index) (Sakurai and Matsui, 1992; Harter, 1979). As non-cognitive ability measures, we created the RSES and CPCS indexes based on these questions.[11]

To assess the possible long-term impact of the intervention, we also collected students' results from the PSC examination, a nationally administered primary education completion test by the Ministry of Primary and Mass Education.[12] We particularly focus on PSC math results, given that our intervention was related to math problem-solving skills. Grade-four (grade-three) students had a chance to take the PSC exam for about 8 months (20 months) after the end of the intervention in December 2016 (December 2017).[13]

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

2.5 Balancing Test Results

Baseline balance tests are performed by comparing the main variables of interest between the treatment and control group students: DT scores per min, DT score, DT time, PTSII-C score, and variables measuring non-cognitive abilities (RSES Index, CPCS Index). The mean and standard deviation of all raw scores and student characteristics are reported in Table 1. As shown, there are no significant differences in average baseline scores between the treatment and control school students (baseline balance), suggesting the success of randomization.

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

[12] Those who wish to pursue further education need to pass this exam, and based on the exam results, letter grades from A+ to A, A-, B, C, D, and F are assigned: if the score is in the range of 80 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. The subjects include math and English in addition to other subjects ⟨ `http://www.educationboard.gov.bd/computer/grading_system.php` ⟩

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

Table 1. Summary Statistics and Baseline Balance

Dependent Variable	Treatment	Control	Difference	N
DT Score per min[a]	4.829	4.619	0.210	825
	[1.898]	[1.767]	(0.309)	
DT Score	46.979	45.958	1.021	825
	[15.653]	[17.357]	(2.817)	
DT Time	9.907	9.967	-0.060	825
	[0.964]	[0.269]	(0.079)	
PTSII-C Score[b]	34.567	38.767	-4.200	905
	[10.308]	[15.251]	(3.400)	
RSES Index[c]	-0.000	-0.003	0.003	812
	[0.392]	[0.452]	(0.062)	
CPCS Index[c]	0.030	-0.045	0.075	812
	[0.370]	[0.426]	(0.053)	
Female	0.580	0.621	-0.041	974
	[0.494]	[0.486]	(0.025)	
Age	9.936	9.912	0.024	974
	[1.099]	[1.193]	(0.294)	

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

Initially, there were 974 students from 34 schools in our sample. However, we work with a different sample of students for different outcome measures for some obvious reasons. First, students from 5 schools were dropped from the sample due to mismanagement (e.g., offering harder/easier diagnostic tests) in conducting baseline DT, which resulted in 825 students for the DT score analysis. Second, some students missed the baseline PTSII tests, which resulted in a sample of 905 students for the PTSII-C analysis. In addition, some students could not answer all the non-cognitive survey questions due to time constraints; therefore, the sample size for the RSES Index and CPCS Index was smaller (812) than the PTSII-C sample.

2.6 Sample Attrition

While there has been some attrition in our sample at the endline, the attrition rate is not significantly different between the treatment and control groups, except in the case of PTSII-C outcome (Table C1 in Appendix C). However, we do not find any significant differences in baseline PTSII-C scores among the attrition sample (treatment vs. control). Accordingly, these results suggest that attrition will not bias our impact estimates.

3 Empirical Specification and Results

We employ the canonical difference-in-differences model to estimate the impact of the Kumon intervention on the measures of cognitive and 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 taking 1 for endline and 0 for baseline; and u and ε are student fixed effects and the error term, respectively. The average treatment effects on the treated can be captured by the estimated δ. For the estimation, we take the first difference of the original level equation, whereby the dependent variable captures improvements in cognitive or 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).[14]

To investigate heterogeneous treatment effects, we estimate 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); and 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 values of the respective outcome measures at the baseline.[15] The parameters of interest are δ for different initial ability types.

3.1 Impacts on Cognitive and Non-cognitive Abilities

The first four columns in Panel A of Table 2 report the results of estimating the equation (1), using standardized cognitive outcomes, 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 min and PTSII-C scores. The magnitude of the impact is sizable: a 2.073 standard deviation in terms of DT scores per min. While this effect size may seem surprisingly high compared to the effect size of education interventions elsewhere, it should be noted here that the effect size on DT score per min is due to a substantial reduction in test completion time measured as DT time (-2.122 s.d.). However, the effect size of the DT score (0.501 s.d.), that is, improvement in the raw test score, is consistent with previous findings in the literature wherein it is found to be effective in improving learning outcomes. Unlike previous studies that have used test scores to determine cognitive ability, we use test scores per min (DT score per min), as our intervention is designed to increase students' abilities to solve math problems in a time-efficient manner, an important ability in pursuing more complex materials in higher education.

[14] 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.).

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

Table 2. Impact of Kumon on Students' Learning Outcomes

Dependent Variable	DT Score per min[a] (1)	DT Score (2)	DT Time (3)	PTSII-C Score[b] (4)	RSES Index[c] (5)	CPCS Index[c] (6)
Panel A: First Difference Estimates						
Treatment	2.073***	0.501**	-2.122***	1.212***	0.026	-0.095
	(0.570)	(0.226)	(0.544)	(0.292)	(0.185)	(0.173)
Constant	0.839	0.521	-0.881	0.679	0.067	0.148
	(0.158)	(0.142)	(0.227)	(0.212)	(0.108)	(0.112)
Num of Obs.	663	663	663	787	696	696
R-squared	0.168	0.048	0.182	0.193	0.000	0.001
p-value (individual hypothesis testing)	0.001	0.035	0.001	0.000	0.891	0.588
p-value (individual hypothesis testing, wild bootstrap)	0.000	0.028	0.002	0.000	0.889	0.593
p-value (Romano-Wolf stepdown p-value)	0.000	0.000	0.000	0.000	0.851	0.485
Panel B: Endline Estimates						
Treatment	2.103***	0.490***	-2.203***	0.925***	0.120	0.179
	(0.548)	(0.137)	(0.552)	(0.212)	(0.160)	(0.149)
Constant	0.831	0.600	-0.722	0.859	-0.010	-0.031
	(0.133)	(0.104)	(0.226)	(0.124)	(0.099)	(0.089)
Num of Obs.	663	663	663	787	696	696
R-squared	0.182	0.095	0.211	0.152	0.003	0.007
p-value (individual hypothesis testing)	0.001	0.001	0.000	0.000	0.458	0.241
p-value (individual hypothesis testing, wild bootstrap)	0.000	0.000	0.002	0.000	0.476	0.266
p-value (Romano-Wolf stepdown p-value)	0.000	0.000	0.000	0.000	0.248	0.020

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

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

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

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

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

	Initial RSES Index					Initial CPCS Index				
	Initial DT Score			Initial PTSII-C Score		Initial DT Score			Initial PTSII-C Score	
Dependent Variable	DT Score per min[a] (1)	DT Score (2)	DT Time (3)	PTSII-C Score[b] (4)	RSES Index[c] (5)	DT Score per min[a] (6)	DT Score (7)	DT Time (8)	PTSII-C Score[b] (9)	CPCS Index[c] (10)
Panel A: High Initial Cognitive and High Initial Non-cognitive Group[d]										
Treatment	2.830***	0.347	-2.968***	1.327**	-0.029	2.961***	0.360	-3.019***	1.221*	0.083
	(0.810)	(0.218)	(0.669)	(0.401)	(0.217)	(0.863)	(0.205)	(0.706)	(0.442)	(0.238)
Constant	0.167	-0.092	-0.549	0.067	-0.506	0.344	-0.058	-0.842	0.066	-0.505
	(0.158)	(0.115)	(0.201)	(0.269)	(0.167)	(0.228)	(0.114)	(0.308)	(0.294)	(0.177)
Num of Obs.	147	147	147	179	179	145	145	145	184	184
R-squared	0.255	0.059	0.285	0.245	0.000	0.260	0.066	0.284	0.199	0.001
Panel B: High Initial Cognitive and Low Initial Non-cognitive Group[d]										
Treatment	3.073***	0.312	-3.707***	1.548***	-0.012	2.686***	0.262	-3.361***	1.716**	-0.161
	(0.894)	(0.190)	(0.733)	(0.404)	(0.270)	(0.847)	(0.218)	(0.813)	(0.325)	(0.282)
Constant	0.418	0.035	-0.793	-0.017	0.934	0.252	0.001	-0.521	-0.015	0.907
	(0.209)	(0.115)	(0.335)	(0.270)	(0.182)	(0.098)	(0.103)	(0.229)	(0.224)	(0.207)
Num of obs.	123	123	123	160	160	125	125	125	155	155
R-squared	0.314	0.049	0.452	0.255	0.000	0.278	0.032	0.407	0.320	0.004
Panel C: Low Initial Cognitive and High Initial Non-cognitive Group[d]										
Treatment	2.454***	0.611**	-2.481***	1.194***	0.156	2.246***	0.564**	-2.298***	1.060**	0.216
	(0.961)	(0.243)	(0.820)	(0.283)	(0.176)	(0.941)	(0.239)	(0.830)	(0.286)	(0.211)
Constant	1.083	1.027	-0.580	1.034	-0.767	1.132	1.073	-0.552	1.144	-0.789
	(0.145)	(0.185)	(0.262)	(0.194)	(0.093)	(0.131)	(0.128)	(0.228)	(0.225)	(0.145)
Num of Obs.	125	125	125	166	166	127	127	127	164	164
R-squared	0.160	0.093	0.198	0.253	0.004	0.132	0.070	0.167	0.199	0.008
Panel D: Low Initial Cognitive and Low Initial Non-cognitive Group[d]										
Treatment	1.127***	0.440	-1.433**	0.928***	0.340*	1.311***	0.514*	-1.610***	1.044***	0.272
	(0.379)	(0.264)	(0.490)	(0.215)	(0.176)	(0.379)	(0.230)	(0.511)	(0.238)	(0.204)
Constant	1.704	1.409	-0.863	1.292	0.406	1.639	1.360	-0.863	1.204	0.539
	(0.255)	(0.168)	(0.260)	(0.136)	(0.062)	(0.259)	(0.189)	(0.303)	(0.124)	(0.068)
Num of obs.	149	149	149	191	191	147	147	147	193	193
R-squared	0.089	0.049	0.124	0.196	0.021	0.117	0.071	0.151	0.232	0.012

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

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

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

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

[d]: The Initial Cognitive Score stands for the baseline DT Score for columns (1)-(3) as well as (6)-(8) and the baseline PTSII-C Score for columns (4),(5),(9), and (10). The Initial Non-cognitive Score stands for the baseline RSES Index for columns (1)-(5). The baseline CPCS Index is used in columns (6)-(10).

We also employ an alternative measure of cognitive ability, PTSII-C, to estimate Equation (1). As can be seen, the estimated effect size using PTSII-C is a 1.212 standard deviation. In summary, as for the impacts measured by DT score per min, the time-saving channel appears to be critical, while PTSII-C captures the impacts on arithmetic skills themselves, in which we find larger impacts. In contrast, regarding the non-cognitive outcomes reported in the last two columns in Panel A of Table 2, the homogeneous treatment effect size estimates are insignificant. While several hypotheses are tested simultaneously, the results are qualitatively the same even when we correct for multiple hypothesis testing, using the Romano–Wolf procedure (Romano and Wolf, 2005).[16] In Panel B of Table 2, we confirm these findings with endline data only.[17]

The heterogeneous treatment effects are reported in Panels A through D of Table 3. We find positive and significant coefficients of cognitive outcomes for all four initial ability types. The magnitudes of DT score per min are larger for students with high-initial cognitive abilities (high-high type and high-low type), while they are smallest for the students with low initial abilities in both measures (low-low type). Regarding non-cognitive outcomes, however, we find suggestive evidence of the catch-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 Index), while others do not show significant effects in non-cognitive scores.[18]

These results support a "building block" story of non-cognitive ability: the Kumon intervention first improves the non-cognitive ability of those who are initially lagging in both cognitive and non-cognitive abilities (i.e., catch-up on non-cognitive ability for low-low type); in turn, it improves the cognitive ability of those with sufficiently improved non-cognitive ability (i.e., higher impacts on cognitive ability compared to low-high type to low-low type).

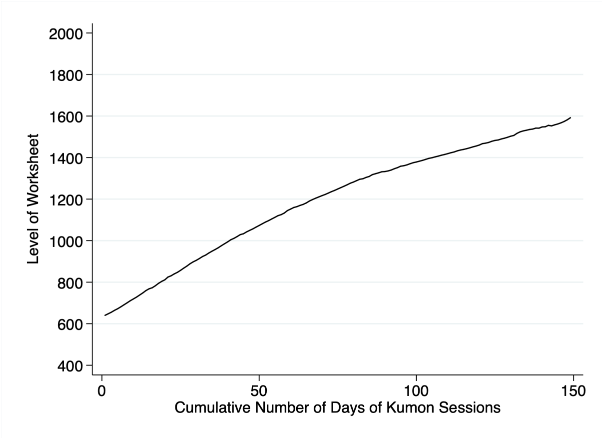
Since students in the treatment schools have studied Kumon materials for an additional 30 minutes per day, one might argue that the impact estimates we present here may be attributed to longer session times in schools and not merely due to the Kumon intervention. We investigate this possibility. By attending the Kumon session, treatment group students study 50 percent longer in math class than control group students (who only attend the regular BRAC math class for one hour). According to column (1) in Panel A of Table 3, 0.839 s.d. is

[16] We report three types of p-values in Table 2. First, we calculate the p-value (individual hypothesis testing) by running each regression separately with school-level clustering. Next, p-value (individual hypothesis testing, wild bootstrap) is calculated by running each regression separately with school-level clustering using the wild bootstrap. Lastly, the p-value (Romano–Wolf stepdown p-value) is based on multiple hypothesis testing with school-level clustering.

[17] As a robustness check, we report the results using the student sample with records of all test results in baseline and endline. As shown in Table D1 of Appendix D, the impact estimates are qualitatively the same.

[18] For the low-low type students, we further examine these heterogenous treatment effects nonparametrically by comparing the cumulative density functions of the first-differences of the non-cognitive scores between the treatment and control groups. Based on the asymptotic p-value of the Kolmogorov–Smirnov test, we reject the null hypothesis that the two distributions are the same for the RSES measure (Figure E3). We can also see the clear tendency of stochastic dominance with the CPCS measure, while the null hypothesis cannot be rejected (Figure E4). Figures E1 and E2 show the cumulative density functions of the baseline non-cognitive scores measured by the RSES and CPCS indexes for the treatment and control groups, respectively. Based on the Kolmogorov–Smirnov test, we cannot reject the null hypothesis of the same distributions; therefore, the baseline is balanced between the treatment group and the control group.

Figure 1. Average Learning Curve with Kumon Worksheets



Note: The level of worksheets is converted into an integer.

the improvement in DT per min for the control group by attending the regular BRAC math class only. If the impact of extending the math learning hours is linear, 50 percent longer hours of learning math should be equivalent to 1.259 s.d. (=0.839 s.d.$\times$1.5) worth of impacts measured in DT per min. If we subtract this longer study-hour effect size (1.259 s.d.) from the treatment coefficient (2.073 s.d.), we have 0.8145 s.d. This is still a fairly large treatment effect. [19] In fact, as Figure 1 shows, the Kumon learning curve is slightly concave, indicating that the students' rate of improvements in math learning outcomes decreases as the study hours become longer. Based on this observation, the back-of-the-envelope counterfactual calculation of longer study hours using the linear assumption might be conservative. Furthermore, we exploit the fact that some treatment schools conducted Kumon sessions for at least five minutes longer. Using these time variations in the Kumon sessions, we examine the impact of the longer study time of Kumon (Table 4). 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. An additional five minutes did not change the treatment effects, which may be due to the flattening learning curve (sharply decreasing marginal impact beyond 30 minutes).

[19]Similarly, if we use the effect size of PTSII-C (1.212 s.d.) and subtract 50 percent longer study-hour effect size (0.679 s.d.$\times$1.5), we have 0.194 s.d..

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

Dependent Variable	DT Score per min[a] (1)	DT Score (2)	DT Time (3)	PTSII-C Score[b] (4)	RSES Index[c] (5)	CPCS Index[c] (6)
Treatment	2.321***	0.350	-2.582***	1.103**	-0.031	-0.136
	(0.826)	(0.246)	(0.679)	(0.360)	(0.224)	(0.192)
Treatment × Longer Session	-0.716	0.437	1.331	0.311	0.177	0.132
	(0.848)	(0.344)	(0.805)	(0.334)	(0.290)	(0.288)
Constant	0.839***	0.521***	-0.881***	0.679***	0.067	0.148
	(0.158)	(0.142)	(0.227)	(0.212)	(0.108)	(0.112)
Num of Obs.	663	663	663	787	696	696
R-squared	0.176	0.063	0.210	0.199	0.002	0.002

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

3.2 Long-term Impact

To assess the long-term impact of the intervention, we use additional information from national examination achievements after 8 months and 20 months of the intervention for the fourth and third grade students in our sample, respectively. Specifically, we use information about the PSC examination take-up and dropouts as well as math scores obtained by students in our sample.[20] We found that the PSC take-up rate is not significantly different between treatment and control school students.[21] Importantly, we must address the potential selection bias when comparing the improvements in cognitive ability using PSC math scores and our measure of math ability. 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. However, the average baseline PTSII-C score is not statistically different between the treatment and control school students who took the PSC exam.[22]

To assess the long-term impact, we use the difference in standardized scores from the PSC math exam and the baseline PTSII-C test as the dependent variable in a difference-in-differences framework that controls for individual fixed effects. Furthermore, to mitigate potential selection bias arising from the de-

[20]We collected students' PSC registration IDs from the BPS branch offices and the teachers at 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).

[21]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 stopping education 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 that 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.

[22]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 (p-value: 0.088) at 10 percent significance level. Similarly, the mean of PTS-C scores among PSC takes at the treatment schools is -0.105, while that of the control schools is 0.334, which is not statistically significantly different, but the difference is 0.439 (p-value: 0.110) at the 1 percent significance level.

cision of taking the PSC examination, we employ propensity score matching (PSM), whereby we match the sample based on pre-treatment student characteristics (i.e., student age, age squared, and gender). As shown in Table 5, the results suggest that students from the treatment schools received 0.25 s.d. higher scores than those from control schools (Table 5).[23] Overall, we find a modest long-term impact of the intervention on cognitive ability.

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

Dependent Variable	First Difference of PSC Math Score and Baseline PTSII-C Score			
	PSM regression		IPW regression	
	ATT estimates (1)	ATE estimates (2)	ATT estimates (3)	ATE estimates (4)
Treatment	0.248**	0.247**	0.258**	0.249**
	(0.124)	(0.120)	(0.121)	(0.118)
Constant			-0.266***	-0.238***
			(0.093)	(0.089)
Num of Obs.	443	443	443	443

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

3.3 Teacher Assessment Ability

In addition to student outcomes, we examine the impact of the intervention on teachers' abilities to assess their students' performance. We hypothesize that teachers may be able to improve their own understanding and assessment of student' abilities, as the intervention will allow them to gain more information about students' abilities from the daily progress records. Using the absolute distance between teachers' assessment scores and students' test scores (for each student) as a dependent variable, we conduct the DID analysis. As shown in Table 6, we find a significant improvement in teachers' abilities to assess students' performance in both DT and PTSII-C scores (i.e., a negative sign indicates that the assessment scale is closer to the actual test score scale).

These positive impacts on BPS teachers are unintended but unsurprising, 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 obtain a precise idea of each student's ability. While this may suggest that teachers could have modified their teaching in program schools, we find no significant difference in teaching hours or home workloads between treatment and control schools. We agree that better information about students' progress gives teachers in treatment schools the ability to more accurately assess students' abilities. The Kumon learning approach has good potential for reducing teachers' stereotyping of students by providing them with better information about their students.

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

Table 6. Association between Teachers' Assessment and Student Performance

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

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

4 Comparing Benefits and Costs

Following Duflo (2001) and Heckman et al. (2010), we calculate the benefit-cost ratio (B-C ratio) and the internal rate of return (IRR). Regarding benefits, we use our long-term impact estimate on math PSC scores (Table 5) and estimated wage returns to numeracy skills from Nordman, Sarr and Sharma (2015) that use the matched employer-employee data. 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.[24] We assume that the benefit will last from 1 year to 44 years, considering working age as 16 to 59 and an annual discount rate of five percent, following Duflo (2001). The deadweight loss factor is unused because this program did not involve tax spending or revenue.

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

[24]The first estimate is taken from our results on the PSC exam, and we use the most conservative number (PSM-ATE estimates), 0.247, in Table 5. 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 of Nordman, Sarr and Sharma (2015) (50.91), multiplied by 40 hours per week and 52 weeks. Then, we calculate the life cycle profile of earnings based on the estimates of the returns to tenure and tenure-squared in Nordman, Sarr and Sharma (2015)'s regression (0.037 and -0.00067).

Taka or 113 (124) USD for eight months.

Under the minimum (maximum) cost assumption, the benefit-cost ratio exceeds one when the benefits last for more than fifteen (more than eighteen) years, as shown in Figure F1 (Figure F2) in Appendix F. 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 nine (ten) years with the minimum (maximum) cost (Figures F1 and F2).

5 Conclusions

In this study, 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 by supplementing the regular curriculum. Specifically, we have implemented a field experiment to test the effectiveness of the Kumon mathematics learning program in 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 that provides students with a suitable amount of mental stimulus to enhance their academic and self-learning abilities.

After eight months of the intervention, we find significant and robust improvements in students' cognitive abilities measured by two mathematics tests. The magnitude of the effect size range from 0.501 to 1.212. These impacts on cognitive ability are consistent with some highly effective existing interventions, such as the 0.75 standard deviation 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 pupils with initially low non-cognitive and cognitive abilities. Furthermore, we have demonstrated the long-term impact of the intervention as compared with students' achievements on the national-level examination taken eight and twenty 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 may have benefited from the Kumon intervention by gaining more objective information about students' skill levels.

Our study contributes to the existing literature on TaRL and self-learning at the right level by showing the effectiveness of a paper-and-pencil-based self-learning program, which is not constrained by inadequate ICT access, a barrier that is often encountered in developing countries. Our findings may be generalizable in similar socioeconomic environments given that Kumon has already been market-tested and extended globally. We believe Kumon could be a cost-effective complementary intervention to existing lecture-style primary education. Given its focus, the current paper does not detail the mechanisms behind the impact of the Kumon method. Uncovering these mechanisms would be an important task for future research. In a companion paper, for example, we investigate the peer effects on classroom learning among treatment students (Kawarazaki et al., 2020).

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

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

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

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2008. "Bootstrap-based Improvements for Inference with Clustered Errors." *Review of Economics and Statistics*, 90(3): 414–427.

Chowdhury, Ahmed Mushtaque Raza, Andrew Jenkins, and Marziana Mahfuz Nandita. 2014. "Measuring the Effects of Interventions in BRAC, and How This Has Driven 'Development'." *Journal of Development Effectiveness*, 6(4): 407–424.

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

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

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

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

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

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

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

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

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

Heckman, James J, Jora Stixrud, and Sergio Urzua. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *Journal of Labor Economics*, 24(3): 411–482.

Heckman, James J., Seong Hyeok Moon, Rodrigo Pinto, Peter A. Savelyev, and Adam Yavitz. 2010. "The rate of return to the HighScope Perry Preschool Program." *Journal of Public Economics*, 94(1-2): 114–128.

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

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.

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

Nath, Samir Ranjan. 2012. "Competencies achievement of BRAC School Students: Trends, comparisons and predictors." *Research monograph series no. 51, Dhaka: BRAC.*

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

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

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

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

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.

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

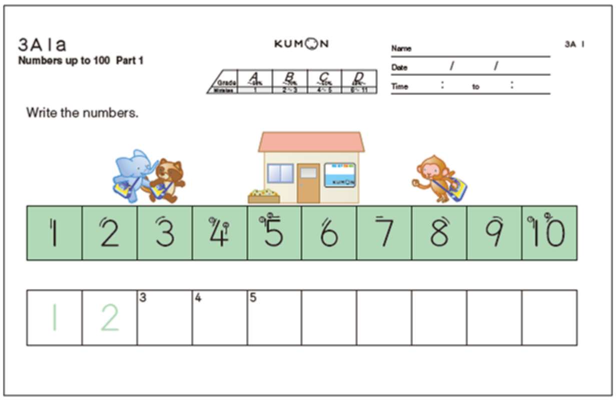
Watkins, Kevin. 2000. *The Oxfam Education Report.* Oxford: Oxfam International.

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

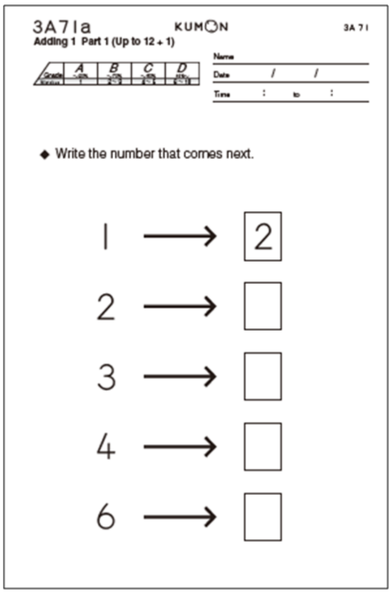
Online Appendix

A Kumon Method Worksheet Examples

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

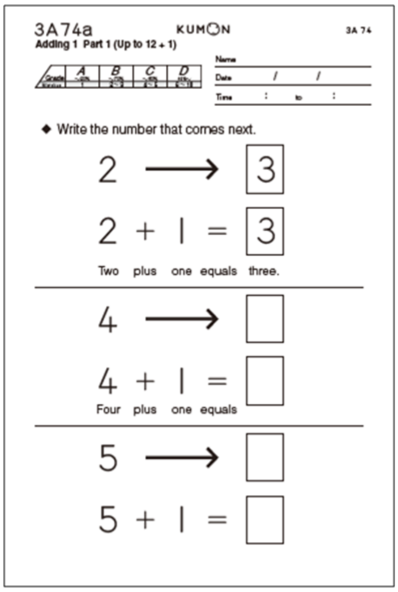
1 → 2

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.

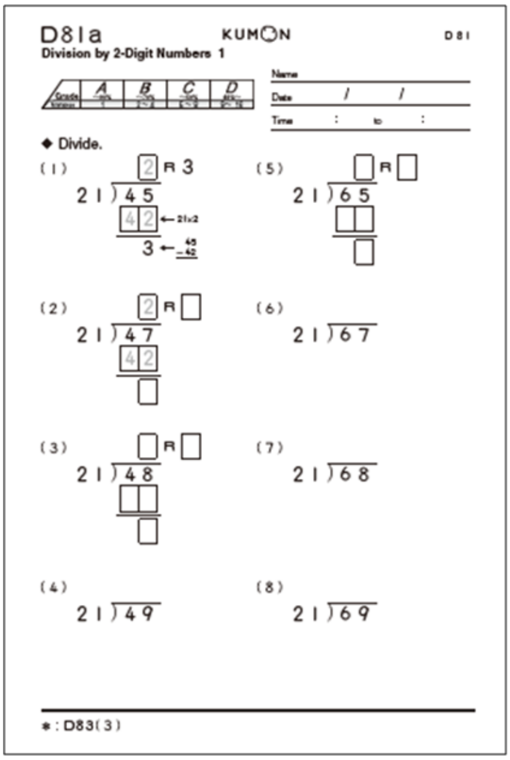
4 → ☐

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) 21) 45 = 2 R 3; 42 ← 21×2; 3 ← 45 − 42

(2) 21) 47 = 2 R ☐; 42

(3) 21) 48 = ☐ R ☐

(4) 21) 49

(5) 21) 65 = ☐ R ☐

(6) 21) 67

(7) 21) 68

(8) 21) 69

＊: D83(3)

B Non-cognitive Ability Survey Questions

Table B1. PTS II survey questions for measuring non-cognitive abilities

Number	Question in English	RSES	CPCS
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 in myself.	x	x
11	I am satisfied with myself.	x	x
12	Even if I fail, I think I can get better and better at things if I keep trying.		
13	I like to do calculations.		
14	I can calculate in my head when I go shopping.		
15	I think speed is important when solving problems.		
16	While studying, I believe everything will go well if I correctly follow the instructions.		
17	I am more motivated when people praise me.		
18	I always volunteer in class.		
19	I enjoy studying.		
20	School is fun.		
21	I do things better when I have a goal.		
22	There are many things I want to learn more about.		
23	a. I have a role model around me. b. There is someone who I want to be like.		
24	I always have someone who I can go to for advice when I am having trouble with my studies.		
25	a. There is someone who I do not want to lose against. b. There is someone who I am always competing with.		
26	I always try to do something when things don't go as expected.		
27	It doesn't matter whether I fail in the beginning because I believe that things will eventually work out.		

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

C Sample Attrition

Table C1. Attrition Analysis

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

Panel B: Attrition Only Sample	
Dependent Variable	Baseline PTSII-C Score
Treatment	-0.178
	(0.209)
Constant	-0.132
	(0.182)
Num of Obs.	118
R-squared	0.013

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

D Robustness Analysis

Table D1. Impact of Kumon on Students' Learning Outcomes: Robustness Analysis

Dependent Variable	DT Score per min[a] (1)	DT Score (2)	DT Time (3)	PTSII-C Score[b] (4)	RSES Index (5)	CPCS Index (6)
Panel A: Fisrt Difference						
Treatment	2.652***	0.629**	-2.779***	1.423***	-0.086	-0.154
	(0.708)	(0.251)	(0.616)	(0.379)	(0.195)	(0.189)
Constant	0.818	0.550	-0.703	0.459	0.106	0.180
	(0.183)	(0.157)	(0.220)	(0.205)	(0.113)	(0.117)
Num of Obs.	509	509	509	509	509	509
R-squared	0.239	0.078	0.288	0.241	0.001	0.003
p-value (individual hypothesis testing)	0.001	0.019	0.000	0.001	0.661	0.422
p-value (individual hypothesis testing, wildbootstrap)	0.000	0.034	0.002	0.002	0.661	0.438
p-value (Romano-Wolf stepdown p-values)	0.000	0.000	0.000	0.000	0.446	0.257
Panel B: Endline Estimate						
Treatment	2.467***	0.396**	-2.844***	1.111***	0.065	0.137
	(0.692)	(0.133)	(0.639)	(0.257)	(0.184)	(0.181)
Constant	0.894	0.722	-0.572	0.766	0.025	-0.006
	(0.155)	(0.094)	(0.233)	(0.115)	(0.101)	(0.092)
Num of Obs.	509	509	509	509	509	509
R-squared	0.216	0.072	0.302	0.216	0.001	0.004
p-value (individual hypothesis testing)	0.002	0.006	0.000	0.000	0.726	0.457
p-value (individual hypothesis testing, wildbootstrap)	0.000	0.012	0.002	0.002	0.765	0.607
p-value (Romano-Wolf stepdown p-values)	0.000	0.000	0.000	0.000	0.475	0.168

Notes: The sample includes students with all test results from both baseline and endline and those without suspicion of cheating. Asymptotic standard errors are shown in parentheses and are clustered at the school level. The superscripts, ***, **, *, denote the statistical significance obtained by clustered wild bootstrap-t procedures at the 1 percent, 5 percent, and 10 percent level, respectively.
[a]: DT Score per min stands for math Diagnostic Test scores per minute.
[b]: PTSII-C Score stands for the math proficiency test scores.
[c]: The Proficiency Test of Self Learning is based on 27 survey questions, of which ten are consistent with the Children's Perceived Competence Scale (CPCS Index) and eight with the Rosenberg Self-Esteem Scale (RSES Index). For each of the non-cognitive type questions, see Appendix B.

E Graphical Evidence of the Effects on Non-cognitive Ability for Children with Initially Low Cognitive and Non-Cognitive Abilities

Figure E1. Cumulative Density Functions of Baseline RSES for Students with Initially Low Cognitive and Non-cognitive Skills

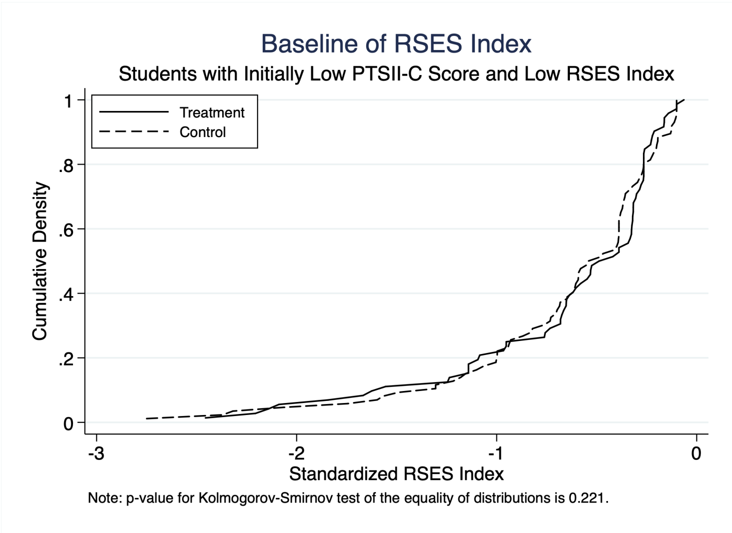


Figure E2. Cumulative Density Functions of Baseline CPCS for Students with Initially Low Cognitive and Non-cognitive Skills

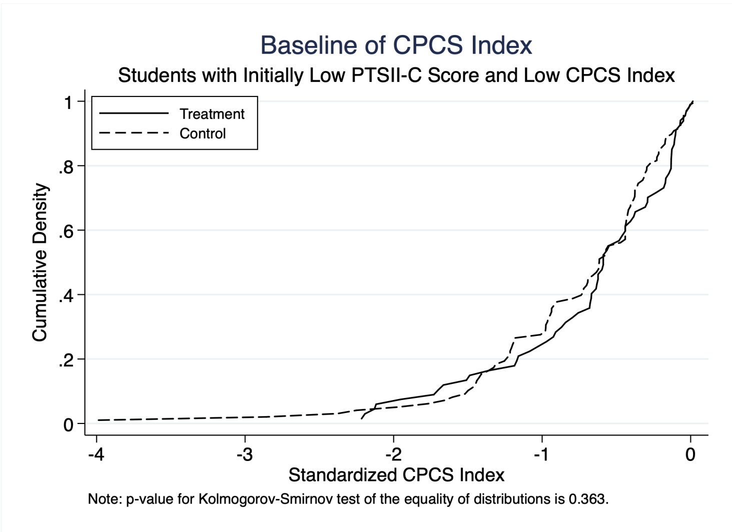


Figure E3. Cumulative Density Functions of the First Difference in RSES for Students with Initially Low Cognitive and Non-cognitive Skills

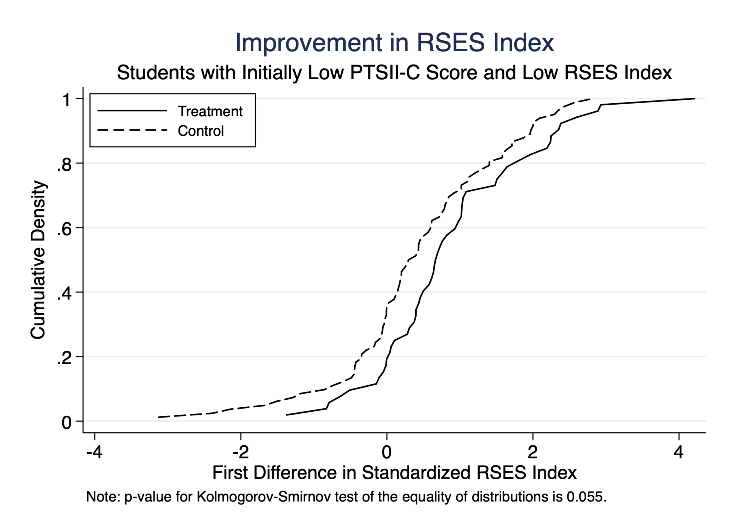
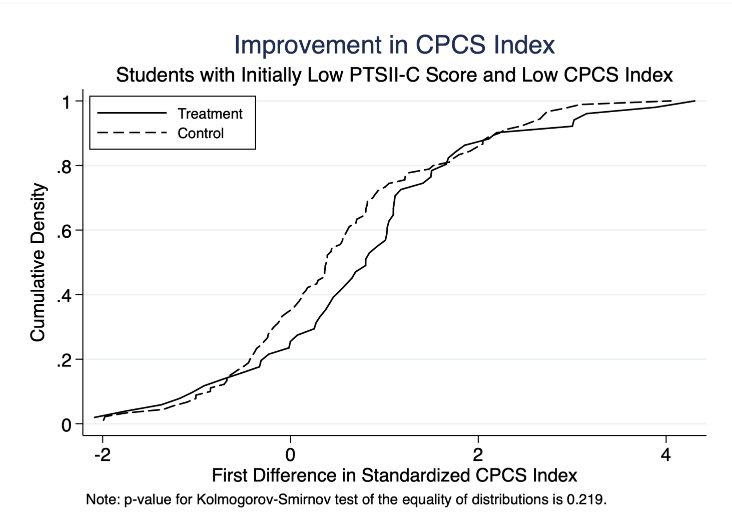
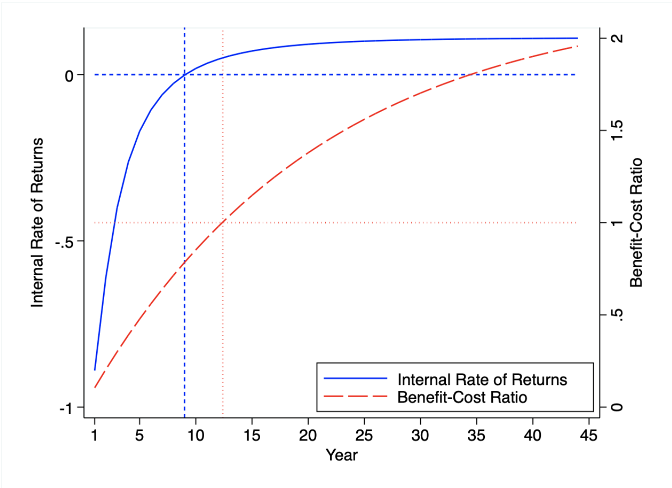


Figure E4. Cumulative Density Functions of the First Difference in CPCS for Students with Initially Low Cognitive and Non-cognitive Skills



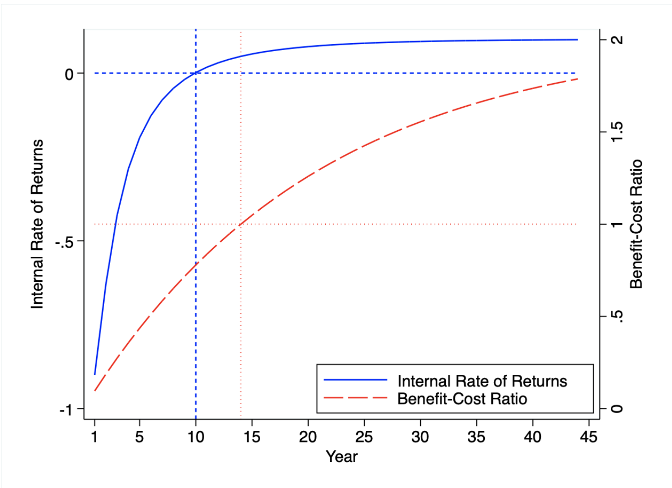
F Graphical Evidence of Benefit-Cost Analysis

Figure F1. Benefit-Cost (B-C) Ratio and Internal Rate of Return (IRR) with Minimum Cost



Notes: The blue solid line indicates the internal rate of return (IRR) and the red long-dashed line indicates the benefit-cost ratio(BC). The blue dashed line indicates IRR = 0 and year = 9, while the red dotted line shows BC = 1 and year = 12.4.

Figure F2. Benefit-Cost (B-C) Ratio and Internal Rate of Return (IRR) with Maximum Cost



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