Individualized Self-learning Program to Improve Primary Education: Evidence from a Randomized Field Experiment in Bangladesh*

[1], Yasuyuki Sawada[†1], Minhaj Mahmud[2], Mai Seki[3], An Le[4], and Hikaru Kawarazaki[5]

[1,4,5] *University of Tokyo*
[2] *Bangladesh Institute of Development Studies*
[1,2,3] *JICA Research Institute*

September 26, 2017

Abstract

This paper reports on the results from a field experiment that tests the effectiveness of the globally popular Kumon learning method in improving the cognitive and non-cognitive abilities of disadvantaged pupils in Bangladesh. Using a randomized control trial design, we study the impact of this individualized self-learning approach among third and fourth graders studying at BRAC non-formal primary schools. The results show that students of both grades in the treatment schools record substantial and significant improvement in their cognitive abilities as measured by two different mathematics tests (Kumon diagnostic test score per minute and proficiency test score) after a period of 8 months, compared to students in the control schools. In terms of non-cognitive abilities, the results give some evidence of positive and significant impacts, particularly on the self-confidence of the pupils. Interestingly, this intervention also had a positive and significant impact on the ability of teachers' to assess their students' performance. Overall our results suggest the wider applicability of a properly designed non-formal education program in solving the learning crisis in developing countries.

Keywords: education, self-learning, cognitive and non-cognitive outcomes, developing countries, randomized control trial

JEL Classification: I20, O12

*We are thankful to Deon Filmer, Dean Karlan, Halsey Rogers, Paul Romer, David Figlio as well as all seminar participants at the North American Summer Meetings of the Econometrics Society 2017, the Asian Meetings of the Econometrics Society 2017, 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 useful comments. This study was prepared by the authors in their own personal capacity. The opinions expressed in this article are the authors' own and do not reflect the view of affiliated organizations.

†Corresponding author: Faculty of Economics, the University of Tokyo

1 Introduction

In many parts of the world there have been tremendous successes achieved under the millennium development goals (MDGs) in terms of school enrolments. According to a UNESCO (2015) estimate, there are 84 million fewer out-of-school children and adolescents now than there were in the year 2000. However, the crisis in the quality of learning remains a serious concern among policy makers. It is observed that significantly fewer children meet a basic level of proficiency in mathematics and reading in developing regions, while 38% of primary school completers in those countries do not have an adequate level of learning outcome (UNESCO, 2014). Given that education is the important link between all the sustainable development goals (SDGs), which set global objectives to end poverty and hunger, improving the quality of education is *sine qua non* for achieving the SDGs. Indeed, Hanushek (2009) concludes that quantitative expansion of schooling without emphasis on quality might be counterproductive for developing countries.

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.[1] In relations to improving learning outcomes, demand side approaches appear to be less promising, compared to supply side interventions such as more teachers and schools (Asim et al., 2016).[2] However, Berlinski and Busso (2015) suggest that the types of pedagogical intervention that require teacher's to adopt high levels of new technology might be counterproductive at least in the short-run. In this context, the pedagogical interventions that match teaching to students learning and ability levels are gaining increasing attention due to their effectiveness in improving learning outcomes. For example, Banerjee et al. (2007) have shown that remedial education programs that teach basic numeracy and literacy skills to children lagging behind in government schools in India have been very effective in enhancing the test scores of children in the treated schools relative to control groups, and that the results persisted a year after the program was implemented. More recently, Banerjee et al. (2016) conducted several randomized evaluations of teaching at the right level in India, and found that both 10-day intensive summer camps as well as school-year interventions using such learning methods were more effective in a setting where students were grouped in terms of initial learning level. Duflo et al. (2011) also found that tracking effectively improves learning outcomes, and suggest that tracking could particularly benefit weaker pupils while addressing stereotyping behaviour on the part of teachers.

In our study, we designed and implemented a randomized control trial (RCT) to address the learning crisis in developing countries (here Bangladesh). We adopt and evaluate a globally popular non-formal individualized education program, the Kumon method of learning (hereafter Kumon), designed to improve both the cognitive and non-cognitive abilities of primary school students in the context of Bangladesh. More precisely, Kumon is a market-tested non-formal education program, which is designed to ensure that students always study at a

[1]These have ranged 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 et al., 2007; Glewwe, 2002).

[2]See Asim et al. (2016) for a meta-analysis of impact evaluation studies focusing to improve learning outcomes in South Asian countries. Other reviews discussing the impacts of interventions on learning outcomes include: Kremer et al. (2013), Ganimian and Murnane (2016), Evans and Popova (2015), McEwan (2015), and Glewwe (2013)

level that is “just right” for them.[3] This philosophy is similar to the “teaching at the right level”, but emphasizes the self-learning aspects of the learning experiences. Students begin at a comfortable staring point and learn new concept in small steps. Self-learning is enforced through hints and examples. Other contributions of this paper are the assessment of the interventions on the non-cognitive abilities of disadvantaged children in developing countries, and the examination of whether these have a complimentary impact on learning outcomes.[4]

Bangladesh is one of the few countries on track to achieve universal primary education. The net primary education enrolment rate (NER) of Bangladesh increased from 62.9% in 2000 to 97.3% in 2013 (Emran, 2014), with a substantial reduction of the gender disparity in access to education (Kono et al., 2016). In addition to dynamic Bangladeshi government policies and pro-active engagement in improving education (Ravallion and Wodon, 2000; Ahmed and Ninno, 2002; Heath and Mobarak, 2015), schooling has been provided by NGOs, coupled with the continued donor support for targeted programs and interventions designed to reach out to the disadvantaged, and these policies played a catalytic role in improving school enrolment.[5] However, despite the glowing achievement of increased school enrolment and the narrowed gender gap, the lack of adequate student learning remain a serious concern in Bangladesh, as in other developing countries. For example, Asadullah and Chaudhury (2013) evaluated the ability of school students to answer simple arithmetic problems, finding a puzzling and imperfect correlation between years of schooling and cognitive outcome.[6] Hence, in order to contribute to policy making designed to remedy the learning crisis in primary education in Bangladesh, our research strategy is to introduce and evaluate the effectiveness of Kumon mathematics module in improving the cognitive and non-cognitive outcomes of disadvantaged young children studying in BRAC Primary Schools (BPS), by implementing a carefully designed randomized control trial study.[7]

In August 2015, Kumon has been introduced for the first time at selected BPS locations in Bangladesh. Out of 179 schools, 34 are randomly selected into treatment and control groups, and roughly 1000 students in these schools are tracked over eight months followed by a baseline study. To preview our findings, after eight months of intervention, Kumon substantially and significantly improved students' cognitive ability, and therefore learning outcomes. This is captured through a series of Kumon Diagnostic Tests (DT) and Proficiency

[3] As of March 2017, there are 4.35 million subject enrolments in 50 countries and regions, according to Kumon Institute of Education.

[4] 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 started shedding new light on the role of non-cognitive abilities such as personality traits, motivations, and preferences (Heckman, 2007, 2006). In fact, recent studies have begun to show that, in explaining education, success in the labor market, or other outcomes, the predictive power of non-cognitive abilities are comparable or exceed that of cognitive skills (Heckman, 2006; James et al., 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.

[5] For example, the largest NGO in Bangladesh, BRAC, operates a non-formal education program, the BRAC Primary School (BPS), which has come to be regarded as one of the most successful interventions in the promotion of education for poor children. BPSs introduced a seasonally adjusted school calendar, which has been a key to their success (Watkins, 2000; Chowdhury et al., 2014).

[6] Their findings suggest that among those who completed primary schooling, only 49 percent could provide 75 percent or more correct answers in a simple arithmetic problem test, 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.

[7] BPS is one of the largest non-formal education programs targeted at disadvantaged populations in Bangladesh, providing a four-year program, which covers the five-year public primary school curriculum. After the final year, Bangladesh government allows BPS students to take the Primary Education Terminal Examination, which is necessary to advance further schooling ⟨http://brac.net/education-programme/item/761-brac-primary-schools⟩

Tests of Self-learning skills (PTSII). The magnitude of the impact ranges from 1.50 - 2.06 standard deviation for grade three, and 2.12 - 2.64 for grade four, both measured by DT test score per minute. These effects are large compared to some existing interventions.[8] While most studies use test scores, we use score per minute in the case of diagnostic tests as our intervention is designed to increase a student's ability to solve math problems in a time efficient manner. In the case of PTSII, the magnitude of the impact is found in the range of 0.78 for grade three, and 1.02 - 1.22 for grade four students respectively. In terms of non-cognitive abilities, the results show some evidence of positive and significant impacts, particularly on the self-confidence of the pupils. Lastly, this intervention also had some positive and significant impacts on a teachers' ability to assess their students' performance. This latter result may suggest that short term intervention becames effective in mitigating teachers' stereotypes by facilitating a better match between teaching and student level, and thereby improving student abilities.

The rest of paper is organized as follows. In Section 2, we outline our experimental design, followed by a description of our data and our baseline test results. Section 3 gives the econometric evaluation framework, followed by empirical results. Section 4 concludes the paper.

2 Experiment Design, Data, and Balancing Test

2.1 Intervention: the Kumon Method of Learning

Kumon has been attracting a wide attention globally as an effective program to strengthen both cognitive and non-cognitive learning outcomes. In Bangladesh, Kumon has been adopted in selected BPS as a pilot program to improve their third and fourth grade students learning outcomes in mathematics. Kumon aims to enable students to develop advanced academic and self-learning ability by ensuring that children are always studying at a level that is just right for them. Students begin from a comfortable starting point suitable to their ability level, regardless of their age or grade level in formal school. The comfortable starting point is usually set slightly below students' concurrent maximum potential capacity to: i) ensure full understanding of the basic concepts as a firm building block of cognitive ability development, and ii) boost students' motivation to continue to study, which also works for the development of their non-cognitive ability. Kumon's 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.[9] All of these worksheets are carefully designed, starting from simple counting to advanced mathematics, with the level of difficulty increasing in small steps.[10] Some worksheets contain example questions with hints, which help students to acquire step by step problem solving skills by themselves. As a result, students can absorb material beyond their school grade level through self-learning, and advance to studying high school level material at

[8] 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 et al. (2011) found a 0.9 standard deviation impact from the peer effects of tracking for the top quartile students in the primary schools of Kenya.

[9] Kumon also has reading related subjects, such as English, but we concentrate on the explanation of the mathematics program, as it is the subject that BPS has focused on at this time.

[10] Appendix 3 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.

an early age. Importantly, slower learners can spend more time on the basics without being rushed to move on to advanced level of materials beyond their understanding.

2.2 Experimental Design

To identify the causal effects of Kumon on young students' cognitive and non-cognitive abilities, we design and conduct an RCT study. We need a design that will allow us to have adequate statistical power to detect a minimum effect of at least 0.4 standard deviation.[11] Considering that randomization is done at cluster (school/class room) level, we assume intracluster correlation of 0.10, and a statistical significance of less than 0.05 for a two-tail test, which means a sample of about 774 (26 clusters) with a statistical power of 0.80. So that we do not lose statistical power due to attrition or other factors, we choose a cluster size of 34, with average 30 students per cluster, that give us a sample of about 1000 students.

We randomly selected 34 BPS having third and fourth graders from the 179 BPS in Dhaka and surrounding areas, 17 schools received Kumon materials and 17 schools did not receive these materials so that they could serve as treatment and control schools, respectively.[12] The resulting sample breakdown by grades is as follows: 19 (out of 48 schools) for the third grade and 15 (out of 131 schools) for the fourth grade are tracked in our study.[13] In total, our study started tracking roughly 1,000 students in these 34 schools. In the schools, we choose only one of the two class shifts (either morning or afternoon), with an average class size of 30 students. The intervention for the treated school students consists of a 30 minute session on Kumon-study prior to the beginning of their regular lesson. Thus, students in the treatment schools come to school earlier than usual during the experiment periods.[14]

Schools normally runs six days a week except on public holidays and teacher training days. Out intervention lasts for 8 months, from August 2015 to April 2016. For the treatment schools, the Kumon Institute of Education Co., Ltd provides an intervention package that consists of a mathematics materials set, an instructor manual comprising sheets for the BRAC teacher to navigate the Kumon-Method of learning as well as employment of two marking assistants.[15] The materials set consists of i) mathematics worksheets with questions at various difficulty levels; and ii) a grading note book to record every-day progress, including the level of worksheet that a student will work on, the number of repetitions needed before achieving the full score on a worksheet, and the number of worksheets that students will finally complete.[16]

[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 et al. (2011), we hypothesize this minimum detectible effect on cognitive ability for high policy impact.

[12] A stratified randomization at the school-branch level might be 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 the 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.

[13] The treatment schools were not overlapping in terms of grade. In other words, treatment schools include either third or fouth grades with Kumon intervention.

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

[15] BRAC field staff are assigned to assist and follow-up on BPS teachers. Three days of preparatory training for BPS teachers and field staff are held prior to launching the program to familiarize teachers about the concept and procedure of the learning method. In addition, three follow-up training sessions are held during the implementation period.

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

During the Kumon session, neither the BRAC teacher nor the two markers play an active role in teaching. The teacher only monitors students and rarely intervenes except when a student appear to be stuck in problem solving. In such cases, the level of the worksheet would be re-adjusted to suit the student's level. Otherwise, the marking assistants simply grade the worksheets and return the results to students.

Until the session ends, students either move on to a new worksheet once achieving the full score in the previous one, or continue trying to correct wrong answers until achieving the full score within a designed timeframe. The uniqueness of this method is characterized by three features. First, students are assigned to an initial level of Kumon mathematics based on individual performance in a diagnostic test (DT), not on the basis of the grade they are in at school. This allows both students and BPS teachers to understand students' cognitive abilities better, as measured through mathematics skills. In fact, Kumon is deliberately designed to set the initial level lower than the maximum capacity of the student, so that each student can continue self-learning from the beginning.

Secondly, the tracking of students' progress and achievements is used to personalize the adjustments of worksheet difficulty levels. It is new in the regular in-class teaching methods in BPS for teachers to conduct daily quizzes to monitor the understanding and progress of each student. For the BPS teachers (of the treatment schools), these detailed progress reports on the worksheets allows them to obtain more objective information about their students' abilities, and their understanding of the mathematics involved. This information may indirectly improve BPS teachers' instructions in the regular BPS classrooms. Moreover, teachers who are not familiar with Kumon instructions could support students learning properly, as Kumon worksheets are laid out in small steps to enable students to self-learn and there is a determined standard time per worksheet to judge whether students can advance to the next level or should repeat a level. This may be appealing to those schools and regions that suffer from a shortage of experienced and/or high-quality teachers, which is a matter of concern not only in developing countries, but also in remote and/or limited-budget district schools in developed countries. While Kumon cannot completely substitute a regular class-room based education with active instructing teachers, these distinctive features are nonetheless promising in providing complimentary learning experiences to the students through intellectual stimulation. Lastly, non-digital instruction and the materials used in this method could also be versatile in a setting that is digitally constrained, or has limited equipment and/or instructors, in less resourced countries and regions.

2.3 Data Description

For our purposes, we gather not only pre- and post-intervention observations about the cognitive and non-cognitive abilities of students in treatment and control schools, but we also collect scores obtained by these students from in-between quizzes taken to monitor progress. In addition, we conduct a teacher survey, as well as a parents /guardian survey construct a comprehensive evaluation of the learning method. According to our design, only the every-day worksheet progress records, the grading book and progress report, are available exclusively for the treatment school students. The diagnostic test measuring cognitive ability is called the Diagnostic Test (DT), and the proficiency test measuring cognitive and non-cognitive is called the Proficiency Test of Self Learning Test (PTSII). DT measures cognitive math abilities, and keeps records of both the score and the time spent for completing the test. The DT test used for this study require students to answer 70 questions within ten minutes. Hence, for the DT test, we calculate test score per minute (DT per min) to determine the cognitive ability of students.

Treatment school students took two levels of DT tests for precision in measuring the starting level, while control school students took one.[17] The PTSII consists of two sections: the first part involves of 6 groups containing a total of 348 math questions aimed at measuring the different dimensions of Math problem solving skills (PTSII score).[18] The second section consists of 27 survey questions, which aim to capture the non-cognitive abilities of students. The survey questions in the PTSII, which aim to measure non-cognitive abilities, are shown in Appendix 1. Among the 27 survey questions that Kumon has prepared, 10 are consistent with the Children's Perceived Competence Scale (CPCS) (Sakurai, 1992; Harter, 1979), and 8 with the Rosenberg Self-Esteem Scale (RSES) (Rosenberg, 1965), and the Grit Scale (GRIT). The rest are more specific to the Kumon-Method of learning original, with the addition of 3 Bangladesh specific questions.

Tables 1 and 2 present the descriptive statistics of the third grade and the fourth grade students respectively.[19] As the first cut for understanding the basic data features of students' learning outcomes, we present unconditional means of DT score per minutes and PTSII score as well as non-cognitive test scores (RSES, CPCS, GRIT consistent non-cognitive scores) of control and treatment groups with the difference between the two groups, for the baseline and endline. We observe substantial improvements not only among the treated school students' learning outcomes, but also those among the control school students. This is mainly because that the same-level tests have been used from baseline to endline. We would also like to highlight the fact that this intervention period is worth one-grade period of schooling for BPS students regardless of the intervention, and BPS itself has been a successful education program to begin with, but is still seeking further improvements in its pedagogy. This imbalance in the unconditional means of one of the cognitive tests (i.e., DT for the third grade) at the baseline is further investigated and discussed in the next section (Table 3).

As reported in the bottom part of Tables 1 and 2, our control (treatment) group sample consist of around 60 percent males in both grade three and four. We also report on the household mean wealth index of the control and treatment group students of both third and fourth grades.[20] Both gender composition and wealth index are not significantly different

[17]Observations with suspicion of cheating in tests are re-adjusted to define the starting level based on Kumon's judgment, and we use dummy variables for controlling any systematic tendency of cheat. Also, some schools assigned partially wrong level of DT tests, so we use dummy for this type of instruction error as well. One school fully failed to comply with instructions, hence those 30 students had to be dropped from the analysis.

[18]Some schools failed to follow the time instruction at the baseline. We address such instruction error during the empirical analysis.

[19]See Appendix 2 for how the tests and survey results are merged, as well as information on the unbalanced sample.

[20]The wealth index is constructed by extracting principal components based on the following variables: last income drawn (How much was the last income drawn?); last income per member (the ratio of last income and total member in the family); average household monthly income In what range does the household's monthly average income fall?); the housing condition of household (high quality =1 if type of dwelling are at, single house, tin shed single house, tin shed semi-detached house; =0 if Katcha single house, Katcha semi-detached house); land holding (how much land, in decimal, do you own other than your homestead); house ownership (=1 if they own the house, live in it without paying rent, pay rent to live in it =0 and if pay subsidized amount to live in it); water source (=1 if tube well and piped tap water, =0 if deep tube well); toilet facility (=2 if latrine in house are ring slab, pit latrine; =1 if septic latrine, 0 if open latrine); access to gas connection (=1 if yes); access to electricity connection (=1 if yes); nutrition status (weekly frequency of meat consumption, egg consumption milk

between the treatment and control groups for both grades. The attrition rates between the baseline and end line are on average 11.3 percent in treatment schools and 15.6 percent in control schools.[21]

2.4 Balancing Test Results

To indicate the success of randomization, baseline sample characteristics between treatment and control group need to be balanced, so that final impact estimates are valid estimates of the effect of the intervention. Therefore before proceeding to analyze the impact of the intervention, we present the baseline balance test results. Specifically, we compare the treatment and control group students with regard to the main outcome variables of interest: DT score per minutes, PTSII scores, and the variables measuring Non-cognitive abilities. The balance test results are reported in Table 3. Columns (1)-(3) and (7)-(9) of Table 3 show the results of testing the balance without conditioning on any further observable characteristics of students, while the columns (4)-(6) and (10)-(12) of Table 3 gives the results conditioned on branch fixed effects and dummies for measurement errors (suspicion of cheating, misguidance of test time, misguidance of test level). The unconditional balance test shows a significant difference between treatment and control group students particularly within third grade with regards to the DT test score per minute, similar to the descriptive statistics reported in Tables 1 and 2. However, the sample is balanced once branch fixed effects and other household characteristics are controlled for. Therefore the subsequent empirical analysis use similar controls when estimating the impact of intervention on major outcome variables.

3 Empirical Specification and Results

3.1 Students Learning Outcomes

3.1.1 Econometric specification

We employ an ANCOVA model to estimate the impact of the Kumon method on students cognitive (measured by DT test score per minute and PTSII score) as well as non-cognitive abilities (measured by aggregate index consistent with RSES, CPCS, GRIT, based on the 27 survey questions). ANCOVA 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 baseline and hence minimize potential sampling error in the impact estimates.

$$Y_{it} = \alpha_1 + \beta Y_{it-1} + \delta_0 d_i + \epsilon_{it}. \quad (1)$$

The dependent variable, Y, captures the level of cognitive and non-cognitive outcomes

consumption by children in household). This relatively higher wealth index may suggest that BPS seems to have attracted students from relatively well-off families because of its success, notwithstanding the claim that BPS targets only disadvantageous families' children. For a review of admission eligibility criteria in BPS see Afroze (2011).

[21] To calculate attrition rates, we consider a student as Dropout if he/she didn't take either the DT test or PTSII in end line. In treatment schools, 57 out of 478 students and in control schools, 82 out of 526 students did not take either DT test or PTSII at the end line for various reasons (e.g., dropouts, absence on the exam days, switch of schools and so on). Table A2 in Appendix 2 shows the characteristics of dropouts and the sample used in the analysis.

and the Kumon intervention is specified by an indicator variable, d, taking 1 for treatment group and 0 for control group to estimate the treatment effect. [22] Unlike the case in a canonical difference-in-difference analysis, which estimates program impact on the within-sample difference between end line and baseline outcomes, ANCOVA analyses are less sensitive to natural within-person variation in the baseline and end line variables. This maximizes statistical power, particularly when outcomes are not strongly auto-correlated (McKenzie, 2012), as is assumed in this study. The ANCOVA model used cluster robust standard errors at the school level. However, given the relatively smaller number of clusters, we used a wild cluster bootstrap procedure for concluding the statistical significance of parameters, following Cameron et al. (2008). 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 are balanced and the regression error vector to be i.i.d..

We estimate the heterogeneous treatment effect by students initial cognitive ability and initial non-cognitive ability, following the specification (2) of, one by one student characteristics.

$$Y_{it} = \alpha_1 + \beta Y_{it-1} + \delta_0 d_i + \sum_j \delta_j d_i \cdot X_{ij} + \epsilon_{it}, \tag{2}$$

where the dependent variable captures level of cognitive or non-cognitive outcome, the X_{ij} variable denotes student characteristics of interest, the indicator, d, represents the Kumon intervention and the parameter δ shows the impact of Kumon.

We conduct robustness check of treatment estimates by specifying different regression models. These include firstly, cross sectional regression using only the end line outcome variables, whereby the endline difference in outcome variables are regressed on treatment status (Equation 3). This is done because of the concern that missing values in baseline responses in case of non-cognitive survey questions (primarily due to administrative problem) could poetically result in selection bias in case of ANCOVA specification. A major concern specific to the non-cognitive analysis is that there was an administrative problem (i.e., instructions on survey test taking time) in the baseline test that caused many students to fail to complete the survey questions. Figure 1 shows the pattern of missing answers to the non-cognitive ability survey questions for the baseline. Compared to Figure 2, which is the same pattern for the end line survey, there are more missing observations in the baseline (i.e., the “white” areas indicate the missing values). Due to the selection bias arising from the number and types of survey questions that students have answered, we employ a cross-sectional analysis using only the endline observations.

$$Y_{it} = \alpha_1 + \delta_0 d_i + \epsilon_{it}. \tag{3}$$

Secondly, we use canonical difference-in difference specification, whereby a single variable, interaction between treatment group and treatment period dummy, indicates treatment.

$$Y_{it} = \alpha_0 + \alpha_1 T_t + \gamma d_i + \delta T_t \cdot d_i + u_i + \varepsilon_{it}, \tag{4}$$

where T is a time dummy; u and ε are student fixed effects and the error term, respectively.

[22] We assign a dummy variable for missing end line observations.

The average treatment effects on the treated can be captured by estimating delta. For the estimation, we take the first difference of the equation (7), whereby the dependent variable, Y, captures the improvements of cognitive or non-cognitive outcomes:

$$\Delta Y_{it} = \alpha_1 + \delta d_i + \Delta\varepsilon_{it}. \quad (5)$$

For the difference-in -difference specification with heterogeneous treatments and additional controls, we take equation (5) and further add the interaction terms between the treatment dummy (d), and student characteristics (X) as follows:

$$\Delta Y_{it} = \alpha_1 + \delta_0 d_i + \sum_j \delta_j d_i \cdot X_{ij} + \Delta\varepsilon_{it}. \quad (6)$$

3.1.2 Results on Cognitive and Non-Cognitive abilities

Tables 4 and 5 report several important findings emerging from the results obtained by using cognitive and non-cognitive ability measures (in terms of test scores) for third and fourth graders. It should be noted that the measures are standardized, so the magnitudes of the impacts are reported in terms of their standard deviations. First, we find significant improvements in cognitive outcomes measured by DT score per minute and PTSII scores in the case of both grade three and grade four students following our main ANCOVA specification (columns (1), (2), (5), (6) and (9) of Table 4). The magnitude of the impact is 2.06 and 2.12 standard deviations respectively for grade three and grade four in terms of DT score per minute. The effect of PTSII is 0.78, and 1.22 standard deviations respectively for grade three and grade four. These findings on cognitive outcomes are robust given the inclusion of initial student characteristics (the heterogeneous treatment effects specification) as reported in columns (3), (4), (7), (8) and (10) of Table 4. In that case, the effect of DT score per minutes is 1.50 and 2.12 standard deviations for grade three and four respectively. The effect size of PTSII is 0.78 and 1.02 standard deviations for grade three and four. We do not find clear patterns in complementarity between non-cognitive and cognitive ability growth in the heterogeneous treatment results. These findings are robust against alternative empirical specifications such as cross-sectional specification using end line outcomes or the difference in differences, as shown in Table 6, as well as in Table 11.

As for the non-cognitive outcomes that are consistent with the canonical socio-psychological measurements at some degree (i.e., Children's Perceived Competence Scale (CPCS), Rosenberg Self-Esteem Scale (RSES), and the Grit Scale (GRIT), we do not find significant impacts in the homogenous treatment specification or the heterogenous specification in either grades (Table 5). While the same patterns are found in the cross-sectional specification using end line outcomes for aggregate non-cognitive ability indexes of CPCS, RSES, and GRIT (Table 7), we find positive and significant impacts in selected individual survey questions related to self-confidence and their level of agreement to the statements: "I can confidently express my opinion", and "I did well in this test" (Table 8).[23] Among these results, self-confidence variable ("I can confidently express my opinion") shows the catch-up effect measured as the cross-term between the treatment and initial non-cognitive measurement is negative (i.e., higher impact on low initial non-cognitive ability students than for the rest). The findings on

[23] Given that same questions were asked to all students, we analyse the full sample controlling for grade dummies.

the non-cognitive outcomes show a slightly more nuanced picture in the heterogeneous specifications of the difference in differences specification. The aggregate non-cognitive ability indexes of CPCS, RSES, and GRIT show Kumon's positive and significant impacts especially among the fourth graders, once the heterogeneity in treatment effects are considered (Table 12). Moreover, for both the third and fourth graders, we find catch-up effects in the non-cognitive abilities, where the non-cognitive skills have improved more among those with a low-initial non-cognitive ability (i.e., negative coefficients on Treatment x Initial Noncognitive Score in Columns (4)-(6), (10)-(12), (16)-(18) of Table 12). These findings, however, are not robust in respect of the empirical specifications as previously reported.

The interpretation of the cognitive ability (measured by DT score per minute and PTSII score) treatment effects is straightforward. The Kumon method improved the mathematical problem solving skills of BPS students in both grades. While part of the improvement could result from the fact that the treated group students become comfortable taking paper-based math quizzes, the sizable impacts suggest that the self-learning approach most likely contributed to the improvement of their actual mathematics skills (particularly their arithmetic skills). Unlike previous studies that use test scores to determine cognitive ability, we use score per minute for diagnostic testing, as our intervention are designed to increase student ability to solve math problems in a time efficient manner. When we looked at the DT score and the DT time separately, it turns out that the large impacts on DT score per minute largely resutl from the improved math-problem solving speed measured by DT time. Moreover, the magnitude of the effect on PTSII scores are comparable to those found by the past studies that focus on teaching at the right level.

As for the interpretation of the non-cognitive survey question results, the findings are not as robust as the cognitive outcomes. However, some positive impacts found in specific non-cognitive survey questions suggest the positive treatment effects of the Kumon method on non-cognitive abilities. In that self-learning raises self-confidence of young learners. In fact, what separates Kumon from stereotypical shadow education systems, is that their materials are aimed not only at improving the cognitive but also the non-cognitive abilities of students, thus the design can be iteratively modified based on the responses of actual students learning the materials.

Additional robustness checks are done to assess i) the impact of longer Kumon sessions (Table 9); and ii) actual scores (continuous variables) on the initial cognitive and non-cognitive abilities for the heterogeneous impact specification (Table 10). The first check utilizes the fact that some treatment schools are reported to conduct 5 minute longer Kumon session. Using this seemingly exogenous time variation, we try to investigate the separate impact of longer class, independent from the impact of the Kumon method itself. We do not find any significant impact in these schools, and could not conclude that the longer study time do have any impact or the data variation is not enough to detect such impact.[24] Based on this test, we conclude that the treatment effects reported previously are all inclusive of extra study time as a part of the Kumon method. The second test shows that the findings are consistent across the qualitative indicator of initial ability as well as the actual scores in the initial cognitive or non-cognitive ability measures under the specification of heterogeneity.

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

3.2 Teacher Assessment Ability

3.2.1 Econometric specification

In addition to the student outcomes, we also attempt to examine the impact of intervention on the ability of teacher to assess of their students' performance. Although this has not been an intended impact of the intervention, we hypothesize that teachers might try to improve their objective assessment ability of individual student skill levels and understanding, as this will allow them to gain more information about student abilities through the daily progress records, than in the current BPS setting.

One of the measurements we construct to capture the accuracy of teacher assessments is the association between teacher evaluation of student performance and student cognitive ability, as indicated by DT score per minute and PTSII's math test score. We quantify the association between these two values for both the treatment and the control group in different time periods separately by using the following specification:

$$\begin{aligned} Y_{it} &= (a_{11} + \beta_{11} X_{it}) d_i T_t + (a_{01} + \beta_{01} X_{it})(1 - d_i) T_t \\ &\quad + (a_{10} + \beta_{10} X_{it}) d_i (1 - T_t) + (a_{00} + \beta_{00} X_{it})(1 - d_i)(1 - T_t) + u_i + \varepsilon_{it}, \end{aligned} \tag{7}$$

where Y, T, d, u, and ε are defined as in the previous section, and X stands for teacher evaluation of student performance. To check for improvement in teacher assessment ability, we first conduct an F-test under the null hypothesis of no difference at the baseline: $H_0 : \beta_{10} - \beta_{00} = 0$. Then, we test the end line differences with the following null hypothesis: $H_0 : \beta_{11} - \beta_{01} = 0$.

Another measurement we considered is the variance in the difference between the standardized value of teacher evaluations and student's actual math test scores. A reduction in this variance implies that the teacher is able to track student math ability, as measured by the DT test score per minute and PTSII test score, more accurately, thus signifying an improvement in their' assessment ability over time. For this measurement, we first standardize both the teacher evaluations and the actual math test scores, and calculate the school-level variance of the difference between these two values. Then, we employ the difference-in-difference framework shown in equation (4) to test the null hypothesis: $H_0 : \delta = 0$.

3.2.2 Results on Teacher Assessment Ability

Our findings on the improvement in a teacher's ability to assess their students' performance are reported in Tables 13 and 14. Table 13 reports the impact on teacher assessment ability measured by the covariance between the teacher's evaluation of student and student cognitive-test scores before/after and with/without treatments, as specified in equation (7). The signs of covariance are negative and larger as the association is higher because the teacher's evaluation is 1 for the highest and 5 for the lowest. We find a significant improvement in teacher ability with the PTSII Scores (No difference in the baseline between the control and treatment groups, while F-test scores are 5.04 and 5.68 for grade three and four respectively in the end line).

Table 14 reports the changes in teacher assessment as measured by the precision measure taking the variance between the difference in standardized teacher's 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 both DT score per minute and PTSII score, while no grades show significant results. Overall, the findings suggest that teacher assessment ability of student Math skills appears to show some improvement, but the significance level of this varies by grade and the type of test.

4 Conclusions

In this paper we investigated the effectiveness of a noble individualized self-learning method also known as Kumon in overcoming the issue of low quality of teaching and learning in the context of developing countries. Specifically, we implemented a carefully-designed 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 the just-right level of study, so that students are provided with a suitable amount of mental stimulus to enhance their academic and self-learning ability. As an overall impact after eight months of intervention, we found significant and fairly robust improvements in student cognitive abilities. The magnitude of this impact ranges from 1.50 - 2.06 standard deviations for grade three, and 2.12 - 2.64 for grade four, both measured by test score per minute. These impacts on cognitive ability as measured by diagnostic tests are large compared to some 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). However, the magnitude of the impact on proficiency test score is found to be in the range of 0.78 for grade three and 1.02 - 1.22 for grade four respectively, which is comparable to previously mentioned effective education intervantion programs.

As for non-cognitive abilities, measured by the aggregated non-cognitive ability indexes consistent with Rosenberg's self-esteem scale, the Children's Perceived Competence Scale, and the Grit Scale, we do not find on robust impacts as cognitive outcomes. Nevertheless, there was some evidence of positive and significant impacts particularly on the self-confidence of the pupils. Lastly, we found some positive impacts on BPS teachers' capacity to assess their students' performance. These findings imply that the BPS teachers might have benefited from the Kumon intervention by gaining more accurate and objective information of student skill levels, which in turn worked towards mitigating the teachers' stereotypes.

The contributions of this paper is summarized as follows. From the policy perspective, this study show that Kumon could be an effective complementary intervention to the existing lecture style primary education for disadvantaged students (e.g., the dropouts from formal education, and those with low-socioeconomic status). Though the BPS itself is a unique and effective non-formal primary education program, the success of the collaboration between the BPS and Kumon may be applicable to the future collaboration between the formal education and private supplementary study programs. Moreover, the non-digital instructions and materials of the method could also be versatile in a setting that is digitally constrained, or has limited equipment and/or instructors in less resourced countries and regions. In fact, the World Bank is going to feature the learning crisis as one of the major issues in the 2018 World Development Report, and is seeking interventions that may be effective in improving student learning outcomes. They do acknowledge the impressive progress towards the MDG goal of universal primary completion; however, they also point to the fact that schooling by itself has not led to learning in many cases, and that this is a wide-spread issue across not only the low-income, but even the middle-income countries (WB, 2017)[25]. Hence, this study's rigorous analysis could contribute not only to solving Bangladesh's learning crisis, but will also be useful in other developing countries facing similar issues, while keeping the contextual differences in mind.

Our study makes a significant contribution to the literature that uses an experimental approach to improve the quality of primary education in developing countries; especially the literature that examines the effectiveness of pedagogical interventions on student learning outcomes (Duflo et al., 2011; Banerjee et al., 2016, 2007). As one of the effective pedagogical interventions, the Kumon Method of Learning in the BPS setting appears to provide BPS teachers with more accurate information of student abilities and understanding. Another dimension of academic contribution is its impact on the non-cognitive ability literature in the field of education, which is especially scarce on evidence in the experience of developing countries with disadvantaged children. In this dimension, we do not see as robust findings. The impacts on non-cognitive outcomes require further investigation to comprehensively understand all the findings. Also, the usual caveats of RCT-based evaluation of development programs also apply to our study. While the Kumon method has been extended globally, the external validity of our findings might be limited to the extent that the sample we study is not representative even in the context of Bangladesh. Nonetheless, these results may be generalizable to similar socioeconomic and policy environments. Future studies should focus on testing the wider applicability of this method, and on estimating the cost effectiveness of intervention in the context of developing countries.

References

Afroze, R. (2011). Assessment of Social and Emotional Skills of BRAC Staff. BRAC report. Mimeo., Dhaka: BRAC.

Ahmed, A. U. and C. D. Ninno (2002). The Food for Education Program in Bangladesh: An Evaluation of its Impact on Educational Attainment and Food Security. Food Consumption and Nutrition Division Discussion Paper 138, Washington DC: Internatinal Food Policy Research Institute.

Asadullah, M. N. and N. Chaudhury (2013). "Primary Education, School Quality and Learning Outcomes in Rural Bangladesh". *Comparative Education Review (forthcoming)*.

Asim, S., R. S. Chase, A. Dar, and A. Schmillen (2016). Improving Learning Outcomes in South Asia: Findings from a Decade of Impact Evaluations. World Bank Research Observer lkw006, Washington DC: World Bank.

Banerjee, A., R. Banerji, J. Berry, E. Duflo, H. Kannan, S. Mukherji, M. Shotland, and M. Walton (2016). Mainstreaming an Effective Intervention: Evidence from Randomized

[25] The World Bank Co-Directors, Dr. Filmer and Dr. Rogers came to JICA for the WDR 2018 Consultation Meeting. This was attended by researchers from JICA-RI and other institutions who shared the knowledge of JICA and the research results of JICA-RI and other Japanese research. A summary of the event is available at: ⟨https://www.jica.go.jp/jica-ri/news/topics/20161121_01.html⟩

Evaluations of "Teaching at the Right Level" in India. NBER Working Paper 22746, Cambridge, MA: National Bureau of Economic Research.

Banerjee, A. and E. Duflo (2006). "Addressing Absence". *Journal of Economic Perspectives 20*(1), 117–132.

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

Berlinski, S. and M. Busso (2015). Challenges in Educational Reform: An Experiment on Active Learning in Mathematics. IDB Working Paper type IDB-WP-561, Washington: Inter-American Development Bank.

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

Chowdhury, A. M. R., A. Jenkins, and M. M. Nandita (2014). "Measuring the Effects of Interventions in BRAC, and How This Has Driven 'Development'". *Journal of Development Effectiveness 6(4)*, 407–424.

Duflo, E., P. Dupas, and M. Kremer (2011). "Peer Effects, Teacher Incentives, and the Impact of Tracking". *American Economic Review 101(5)*, 1739–1774.

Duflo, E., R. Glennerster, and M. Kremer (2007). "Using Randomization in Development Economics Research: A Toolkit". *Handbook of Development Economics 4*, 3895–3962.

Duflo, E. and M. Kremer (2005). "Use of Randomization in the Evaluation of Development Effectiveness". *Evaluating Development Effectiveness 7*, 205–231.

Emran, M. (2014). Bangladesh primary education. Annual sector performance report, Dhaka: Monitoring and Evaluation Division, Directorate of Primary Education, Government of the Peoples Republic of Bangladesh.

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

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

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

Glewwe, P. (Ed.) (2013). *Education Policy in Developing Countries.* Chicago: University of Chicago Press.

Hanushek, E. A. (2009). "School Policy: Implications of Recent Research for Human Capital-Investments in South Asia and Other Developing Countries". *Education Economics 17(3)*, 291–313.

Harter, S. (1979). *Perceived Competence Scale for Children.* Denver: University of Denver.

Heath, R. and A. M. Mobarak (2015). “Manufacturing Growth and the Lives of Bangladeshi Women”. *Journal of Development Economics 115*, 1–15.

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

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

James, J. H., J. E. Humphries, and T. Kautz (Eds.) (2014). *The Myth of Achievement Tests: The GED and the Role of Character in American Life*. Chicago: University of Chicago Press.

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

Kono, H., Y. Sawada, and A. S. Shonchoy (2016). Primary, Secondary, and Tertiary Education in Bangladesh: Achievements and Challenges. Technical report, Tokyo: JICA Research Institute.

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

Kremer, M., C. Brannen, and R. Glennerster (2013). “The Challenge of Education and Learning in the Developing World”. *Science 340*, 297 300.

Lakshminarayana, R., A. Eble, P. Bhakta, C. Frost, P. Boone, D. Elbourne, and V. Mann (2013). “The Support to Rural Indias Public Education System (STRIPES) Trial: A Cluster Randomised Controlled Trial of Supplementary Teaching, Learning Material and Material Support”. *PLoS ONE 8(7)*(e65775).

McEwan, P. 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, D. (2012). “Beyond Baseline and Follow-up The Case for More T in Experiments”. *Journal of Development Economics 99*(2), 210–221.

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

Ravallion, M. and Q. Wodon (2000). “Does Child Labour Displace Schooling? Evidence on Behavioural Responses to an Enrollment Subsidy”. *Economic Journal 110(462)*, 158–175.

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

Ruthbah, U., A. Rabbani, S. Hossain, and G. Sarwar (2016). “Do extra hours of tutoring payoff? Evaluation of a Community Education Programme in Bangladesh”. *Journal of Development Economics 8*(2), 196–215.

Sakurai, S. (Ed.) (1992). *Shinri Sokutei Shyakudo Syu IV (Psychological measurement scale IV): Jido-you Konnpitensu Shyakudo (Competence scale for children) "Jikokachi (Self-Worth)"*. Tokyo: Saiensu-sha.

UNESCO (2014). Teaching and Learning: Achieving Quality for All. The Education for All Global Monitoring Report ED-2014/GMR/GS/1, Paris: UNESCO.

UNESCO (2015). Education for all 2000-2015: Achievements and Challenges. The education for all global monitoring report, Paris: UNESCO.

Watkins, K. (2000). *The Oxfam Education Report*. Oxford: Oxfam International.

WB (2017). Realizing the Promise of Education for Development: Themes and Emerging Messages. In *World Development Report 2018, Concept Note. Washington DC: World Bank*.

5 Tables

Table 1: Summary Statistics Grade 3

Summary Statistics on Students' Performance Measurement

Variable Names	Baseline				End line			
	Control Mean	Treatment Mean	Difference	Observations	Control Mean	Treatment Mean	Difference	Observations
DT Score per Minute[a]	4.44	5.42	0.97**	562	6.77	12.19	5.41***	473
			(0.49)				(0.82)	
PTSII Score	0.06	-0.05	-0.11	582	1.15	1.88	0.74**	501
			(0.33)				(0.28)	
RSES Non-cognitive Score[b]	-0.02	0.02	0.04	582	0.01	-0.01	-0.02	496
			(0.15)				(0.20)	
CPCS Non-cognitive Score[b]	-0.12	0.11	0.23**	582	-0.06	0.04	0.10	496
			(0.12)				(0.19)	
GRIT Non-cognitive Score[b]	-0.17	0.14	0.31**	582	-0.07	0.06	0.13	493
			(0.14)				(0.18)	
Demographics								
Female	0.39	0.42	0.03	564				
			(0.03)					
Wealth Index	-0.05	0.04	0.09	321				
			(0.18)					

Note[a]: DT Score per Minute stands for Diagnostic (Math) Test Score: 70 questions are to be solved correctly in 10 minutes.
Note[b]: The Proficiency Test of Self Learning first half, which consists of 348 math questions. Among the 27 survey questions that Kumon has prepared, 10 are consistent with the Childrens Perceived Competence Scale (CPCS Non-Cognitive Score), 8 with the Rosenberg Self-Esteem Scale (RSES Non-Cognitive Score), and 3 with the Grit Scale (GRIT Non-Cognitive Score).
For survey questions related to each Non-Cognitive Score, see Appendix 1. The responses are recorded on 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.
Note[c]: The wealth index is constructed, by extracting the principal components based on the following variables: last income drawn (How much was the last income drawn?); last income per member(the ratio of last income and total member in the family); average household monthly income In what range does the households monthly average income fall?); the housing condition of the household (high quality =1 if type of dwelling is, a single house, a tin shed single house, a tin shed semi-detached house; =0 if it is a Katcha single house, Katcha semi-detached house); land holding (how much land do you own other than your homestead); house ownership (=1 if they own the house, live in it without paying rent, pay rent to live in it =0 if pay is subsidized to live in it); water source (=1 if tube well and piped tap water, =0 if deep tube well); toilet facility(=2 if latrine in house is a ring slab, pit latrine; =1 if septic latrine, 0 if open latrine); access to gas connection (=1 if yes); access to electricity connection(=1 if yes); nutrition status (weekly frequency of meat consumption, egg consumption milk consumption by children in the household)

Table 2: Summary Statistics Grade 4

Summary Statistics on Students' Performance Measurement

Variable Names	Baseline				End line			
	Control Mean	Treatment Mean	Difference	Observations	Control Mean	Treatment Mean	Difference	Observations
DT Score per Minute[a]	4.84	4.49	-0.35	406	6.27	10.44	4.17**	338
			(0.38)				(1.54)	
PTSII Score	0.24	-0.24	-0.49	422	0.50	1.53	1.03***	336
			(0.30)				(0.22)	
RSES Non-cognitive Score[b]	0.02	-0.02	-0.04	422	-0.13	0.12	0.24	336
			(0.20)				(0.24)	
CPCS Non-cognitive Score[b]	-0.03	0.03	0.06	422	-0.14	0.16	0.30	336
			(0.21)				(0.24)	
GRIT Non-cognitive Score[b]	-0.12	0.12	0.25*	422	0.05	0.00	-0.05	324
			(0.14)				(0.16)	
Demographics								
Female	0.37	0.42	0.06	414				
			(0.05)					
Wealth Index	0.06	-0.05	-0.11	226				
			(0.27)					

Note[a]: DT Score per Minute stands for Diagnostic (Math) Test Score: 70 questions are to be solved correctly in 10 minutes.
Note[b]: The Proficiency Test of Self Learning first half, which consists of 348 math questions. Among the 27 survey questions that Kumon has prepared, 10 are consistent with the Childrens Perceived Competence Scale (CPCS Non-Cognitive Score), 8 with the Rosenberg Self-Esteem Scale (RSES Non-Cognitive Score), and 3 with the Grit Scale (GRIT Non-Cognitive Score).
For survey questions related to each Non-Cognitive Score, see Appendix 1. The responses are recorded on 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.
Note[c]: The wealth index is constructed, by extracting the principal components based on the following variables: last income drawn (How much was the last income drawn?); last income per member(the ratio of last income and total member in the family); average household monthly income In what range does the households monthly average income fall?); the housing condition of the household (high quality =1 if type of dwelling is, a single house, a tin shed single house, a tin shed semi-detached house; =0 if it is a Katcha single house, Katcha semi-detached house); land holding (how much land do you own other than your homestead); house ownership (=1 if they own the house, live in it without paying rent, pay rent to live in it =0 if pay is subsidized to live in it); water source (=1 if tube well and piped tap water, =0 if deep tube well); toilet facility(=2 if latrine in house is a ring slab, pit latrine; =1 if septic latrine, 0 if open latrine); access to gas connection (=1 if yes); access to electricity connection(=1 if yes); nutrition status (weekly frequency of meat consumption, egg consumption milk consumption by children in the household)

Table 3: Baseline Test Results with and without Control Variables

Balancing Test Results												
	Grade 3						Grade 4					
Dependent Variables	DT per min [a] (1)	PTSII Score (2)	RSES[b] (3)	DT per min [a] (4)	PTSII Score (5)	RSES [b] (6)	DT per min [a] (7)	PTSII Score (8)	RSES [b] (9)	DT per min [a] (10)	PTSII Score (11)	RSES [b] (12)
Treatment	**0.540****	**-0.226**	**0.162**	**0.442**	**-0.286**	**0.276**	**-0.235**	**-0.241**	**-0.178**	**-0.353**	**-0.343**	**0.111**
	(0.164)	(0.239)	(0.202)	(0.230)	(0.563)	(0.250)	(0.185)	(0.197)	(0.158)	(0.420)	(0.366)	(0.835)
Constant	0.0614	0.0759	-0.279	0.0499	-0.290	0.794	0.256**	0.0732	0.421***	-2.117	-1.485	0.605
	(0.378)	(0.424)	(0.115)	(1)	(0.935)	(0.452)	(0.222)	(0.178)	(0.150)	(0.170)	(0.587)	(0.773)
Other control variables	No	No	No	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes
Number of Observations	582	582	582	342	342	342	422	422	422	249	249	249
R-squared	0.068	0.369	0.062	0.171	0.480	0.149	0.086	0.299	0.060	0.257	0.491	0.209

Note: Asymptotic standard errors are shown in parentheses and are clustered at school level (34 clusters). The asterisks reflect significance levels obtained by a clustered wild bootstrap-t procedure. ***, **, * denote at 1 percent, 5 percent, and 10 percent level respectively. Note: All the regressions are controlled for branch fixed effects and dummies for measurement errors (suspicion of cheating, misguidance of test time, misguidance of test level).

Control variables: branch dummies, number of days the head of household get sick, number of members in the household, number of adults in the household, number of members in the household finishing primary education, last income, number of male in the household, frequency discussing subjects with child, frequency discussing lessons with child in the past two weeks, frequency of reading story to child, do you agree that parents should help with a child's study, frequency of eating meat of fish,frequency of drinking milk, main source of household income, average income, availability of electricity, availability of gas connection, type of toilet, source of water, house ownership

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

Note[b]: The Proficiency Test of Self Learning first half, which consists of 348 math questions. Among the 27 survey questions that Kumon has prepared, 10 are consistent with the Childrens Perceived Competence Scale (CPCS Non-Cognitive Score), 8 with the Rosenberg Self-Esteem Scale (RSES Non-Cognitive Score), and 3 with the Grit Scale (GRIT Non-Cognitive Score). For survey questions related to each Non-Cognitive Score, see Appendix 1. The responses are recorded on 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 4: Impact of Kumon program on Student Cognitive Outcome: ANCOVA Estimates

Impacts of KUMON Program on Students' Cognitive Outcomes ANCOVA Specification										
	Grade 3				*Grade 4*				*All grades*	
Dependent Variables	DT per min[a] (1)	PTSII Score (2)	DT per min[a] (3)	PTSII Score (4)	DT per min[a] (5)	PTSII Score (6)	DT per min[a] (7)	PTSII Score (8)	PTSII Score (9)	PTSII Score (10)
Treatment vs. Control	**2.056*****	**0.780****	**1.499*****	**0.775*****	**2.636****	**1.223*****	**2.120****	**1.024*****	**0.954*****	**0.867*****
	(0.288)	(0.276)	(0.321)	(0.245)	(0.962)	(0.206)	(0.789)	(0.225)	(0.192)	(0.170)
Baseline Score	0.353***	0.352**	0.244***	0.339***	0.443**	0.337***	0.504*	0.291***	0.336***	0.305***
	(0.111)	(0.112)	(0.0748)	(0.0878)	(0.188)	(0.0732)	(0.248)	(0.0782)	(0.0792)	(0.0633)
Treatment x Initial Cognitive Score (=1 if above median)			0.948**	0.0321			-0.365	0.267		0.126
			(0.384)	(0.253)			(0.906)	(0.214)		(0.198)
Treatment x Initial Noncognitive Score (=1 if above median)			0.379	-0.0856			1.437	0.173		0.0132
			(0.242)	(0.259)			(1.117)	(0.151)		(0.178)
Constant	0.771***	1.127***	0.753***	1.175***	0.907***	0.421***	0.959***	0.435***	0.531***	0.547***
	(0.141)	(0.126)	(0.169)	(0.164)	(0.235)	(0.0960)	(0.261)	(0.0941)	(0.116)	(0.113)
Number of Observations	473	501	473	501	338	336	338	336	837	837
R-squared	0.279	0.194	0.306	0.196	0.207	0.311	0.242	0.324	0.273	0.275

Note: 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; ***, **, * denote at 1 percent, 5 percent, and 10 percent level respectively.
Note[a]: DT per Minute stands for Diagnostic (Math) Test Score per minute: 70 questions are to be solved correctly in 10 minutes.
Both cognitive and non-cognitive test scores are standardized and used in the regression analysis.
Note[b]: The Initial Cognitive Score stands for the DT Score for columns (6), (8)-(10) and the PTSII Score for column (7).
Note[c]: The Initial Non Cognitive Score stands for the average non-cogntive score based on all 27 questions for columns (6) and (7). For columns (8)-(10), RSES, CPCS, and GRIT are used respectively.

Table 5: Impact of Kumon program on Student Non Cognitive Outcome: ANCOVA Estimates

Impacts of KUMON Program on Students' Non Cognitive Outcomes ANCOVA Specification

Dependent Variables	Grade 3						Grade 4						All grades					
	RSES [a] (1)	CPCS [a] (2)	GRIT[a] (3)	RSES [a] (4)	CPCS[a] (5)	GRIT [a] (6)	RSES [a] (7)	CPCS [a] (8)	GRIT [a] (9)	RSES (10)	CPCS (11)	GRIT (12)	RSES [a] (13)	CPCS [a] (14)	GRIT [a] (15)	RSES [a] (16)	CPCS [a] (17)	GRIT [a] (18)
Treatment vs. Control	**-0.0935**	**0.00226**	**0.0436**	**0.222**	**-0.114**	**-0.0316**	**0.244**	**0.292**	**-0.0446**	**-0.275**	**0.292**	**-0.105**	**0.0473**	**0.125**	**0.0192**	**-0.0752**	**0.0559**	**-0.0525**
	(0.174)	(0.159)	(0.138)	(0.241)	(0.155)	(0.212)	(0.220)	(0.219)	(0.151)	(0.175)	(0.239)	(0.171)	(0.137)	(0.129)	(0.100)	(0.146)	(0.139)	(0.142)
Baseline Score	0.0361	0.0645	0.0533	0.0545	0.0648	0.0263	0.185**	0.148**	0.0600	0.162**	0.139**	0.0767	0.107**	0.102**	0.0577*	0.0977*	0.0943**	0.0532
	(0.0735)	(0.0494)	(0.0344)	(0.0687)	(0.0520)	(0.0475)	(0.0821)	(0.0737)	(0.0443)	(0.0721)	(0.0644)	(0.0491)	(0.0499)	(0.0400)	(0.0285)	(0.0496)	(0.0410)	(0.0335)
Treatment x Initial Cognitive Score (=1 if above median)[b]				0.189	0.205	0.0901				0.0228	-0.0265	0.275*				0.118	0.114	0.160
				(0.153)	(0.151)	(0.160)				(0.224)	(0.228)	(0.152)				(0.132)	(0.135)	(0.114)
Treatment x Initial Noncognitive Score (=1 if above median)[c]				0.219*	0.0714	0.0942				0.0288	0.0204	-0.0812				0.165	0.0545	0.0194
				(0.115)	(0.103)	(0.147)				(0.162)	(0.176)	(0.0901)				(0.102)	(0.0968)	(0.0966)
Constant	0.155	0.109	0.114	0.127	0.0885	0.0965	-0.105	-0.115	0.0767	-0.106	-0.115	0.0766	-0.0129	-0.0317	0.0679	-0.0143	-0.0327	0.0675
	(0.122)	(0.116)	(0.0836)	(0.126)	(0.120)	(0.0812)	(0.112)	(0.0986)	(0.104)	(0.112)	(0.0990)	(0.104)	(0.115)	(0.107)	(0.0912)	(0.114)	(0.107)	(0.0904)
Number of Observations	496	496	493	496	496	493	336	336	324	336	336	324	832	832	817	832	832	817
R-squared	0.033	0.041	0.068	0.043	0.046	0.070	0.042	0.043	0.015	0.043	0.043	0.024	0.028	0.035	0.045	0.033	0.036	0.049

Note: 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;
***, **, * denote at 1 percent, 5 percent, and 10 percent level respectively.
Note[a]: The Proficiency Test of Self Learning first half, which consists of 348 math questions. Among the 27 survey questions that Kumon has prepared, 10 are consistent with the Childrens Perceived Competence Scale (CPCS Non-Cognitive Score), 8 with the Rosenberg Self-Esteem Scale (RSES Non-Cognitive Score), and 3 with the Grit Scale (GRIT Non-Cognitive Score). For survey questions related to each Non-Cognitive Score, see Appendix 1.
The responses are recorded on 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.
Note[b]: The Initial Cognitive Score stands for the DT Score for columns (6), (8)-(10) and the PTSII Score for column (7).
Note[c]: The Initial Non Cognitive Score stands for the average non-cognitive score based on all 27 questions for columns (6) and (7). For columns (8)-(10), RSES, CPCS, and GRIT are used respectively.

Table 6: Impact of Kumon program on Student Cognitive Outcome: End line on Treatment Dummy

Impacts of KUMON Program on Students' Cognitive Outcomes Cross-sectional Specification

Dependent Variables	*Grade 3*				*Grade 4*				*All grades*	
	DT per min [a]	PTSII Score	DT per min [a]	PTSII Score	DT per min [a]	PTSII Score	DT per min [a]	PTSII Score	PTSII Score	PTSII Score
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Treatment vs. Control	**2.189*****	**0.738****	**1.499*****	**0.554***	**2.541****	**1.030*****	**2.041****	**0.715*****	**0.855*****	**0.628*****
	(0.330)	(0.283)	(0.327)	(0.296)	(0.940)	(0.222)	(0.725)	(0.210)	(0.191)	(0.196)
Treatment x Initial Cognitive Score (=1 if above median)[b]			1.114**	0.423			-0.0820	0.654***		0.498**
			(0.434)	(0.303)			(0.818)	(0.188)		(0.225)
Treatment x Initial Noncognitive Score (=1 if above median)[c]			0.369	-0.103			1.286	0.200		0.0116
			(0.279)	(0.248)			(1.108)	(0.152)		(0.173)
Constant	0.723***	1.146***	0.740***	1.242***	0.979***	0.505***	1.004***	0.511***	0.583***	0.616***
	(0.138)	(0.164)	(0.171)	(0.175)	(0.245)	(0.107)	(0.251)	(0.109)	(0.120)	(0.118)
Number of Observations	473	501	473	501	338	336	338	336	837	837
R-squared	0.248	0.099	0.294	0.134	0.185	0.220	0.220	0.269	0.190	0.222

Note: Asymptotic standard errors are shown in parentheses and are clustered at school level (34 clusters). The asterisks reflect the significance levels obtained by a clustered wild bootstrap-t procedure; ***, **, * denote at 1 percent, 5 percent, and 10 percent level respectively.
Note[a]: DT per Minute stands for Diagnostic (Math) Test Score per minute: 70 questions are to be solved correctly in 10 minutes.
Both cognitive and non-cognitive test scores are standardized and used in the regression analysis.
Note[b]: The Initial Cognitive Score stands for the DT Score for columns (6), (8)-(10) and the PTSII Score for column (7).
Note[c]: The Initial Non Cognitive Score stands for the average non-cogntive score based on all 27 questions for columns (6) and (7). For columns (8)-(10), RSES, CPCS, and GRIT are used respectively.

Table 7: Impact of Kumon program on Student Non Cognitive Outcome: End line on Treatment Dummy

Impacts of KUMON Program on Students' Non Cognitive Outcome Cross-sectional Specification

Dependent Variables	*Grade 3*						*Grade 4*						*All grades*					
	RSES [a]	CPCS [a]	GRIT [a]	RSES [a]	CPCS [a]	GRIT [a]	RSES [a]	CPCS [a]	GRIT [a]	RSES [a]	CPCS [a]	GRIT [a]	RSES [a]	CPCS [a]	GRIT [a]	RSES [a]	CPCS [a]	GRIT [a]
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
Treatment vs. Control	**-0.0213**	**0.0955**	**0.126**	**-0.269**	**-0.107**	**-0.0508**	**0.242**	**0.300**	**-0.0494**	**0.148**	**0.225**	**-0.154**	**0.0849**	**0.178**	**0.0562**	**-0.269**	**0.0435**	**-0.0821**
	(0.196)	(0.192)	(0.180)	(0.182)	(0.162)	(0.207)	(0.235)	(0.236)	(0.165)	(0.237)	(0.237)	(0.183)	(0.151)	(0.148)	(0.125)	(0.182)	(0.141)	(0.141)
Treatment x Initial Cognitive Score (=1 if above median) [b]				0.194	0.215	0.100				0.0278	-0.0178	0.288				0.194	0.128	0.165
				(0.149)	(0.146)	(0.153)				(0.238)	(0.247)	(0.168)				(0.149)	(0.133)	(0.111)
Treatment x Initial Noncognitive Score (=1 if above median)[c]				0.225*	0.103	0.131				0.198	0.187	0.0210				0.225*	0.126	0.0915
				(0.115)	(0.0947)	(0.121)				(0.203)	(0.218)	(0.0885)				(0.115)	(0.103)	(0.0852)
Constant	0.00852	-0.0561	-0.0670	0.107	0.0570	0.0807	-0.126	-0.139	0.0490	-0.122	-0.136	0.0531	-0.0560	-0.0846	0.00109	0.107	-0.0523	0.0455
	(0.119)	(0.130)	(0.116)	(0.136)	(0.125)	(0.0871)	(0.119)	(0.100)	(0.0952)	(0.120)	(0.102)	(0.0985)	(0.116)	(0.107)	(0.0921)	(0.136)	(0.108)	(0.0889)
Number of Observations	496	496	493	496	496	493	336	336	324	336	336	324	832	832	817	496	832	817
R-squared	0.000	0.002	0.004	0.038	0.040	0.068	0.013	0.021	0.001	0.019	0.026	0.012	0.002	0.007	0.001	0.038	0.027	0.041

Note: 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;
***, **, * denote at 1 percent, 5 percent, and 10 percent level respectively.
Note[a]: The Proficiency Test of Self Learning first half, which consists of 348 math questions. Among the 27 survey questions that Kumon has prepared, 10 are consistent with the Childrens Perceived Competence Scale (CPCS Non-Cognitive Score), 8 with the Rosenberg Self-Esteem Scale (RSES Non-Cognitive Score), and 3 with the Grit Scale (GRIT Non-Cognitive Score). For survey questions related to each Non-Cognitive Score, see Appendix 1.
The responses are recorded on four-point scale: Strongly Agree, Somewhat Agree, Somewhat Disagree, Strongly Disagree. Both cognitive and non-cognitive test scores are standardized and used in the regression analysis.
Note[b]: The Initial Cognitive Score stands for the DT Score for columns (6), (8)-(10) and the PTSII Score for column (7).
Note[c]: The Initial Non Cognitive Score stands for the average non-cognitive score based on all 27 questions for columns (6) and (7). For columns (8)-(10), RSES, CPCS, and GRIT are used respectively.

Table 8: Impact of Kumon program on Student Outcome: End line on Treatment Dummy Survey Questions

Impacts of KUMON Program Individual Survey Questions								
Dependent Variables	Self-confidence (1) [a]				Self-confidence (2) [b]			
	End line	End line	ANCOVA	ANCOVA	End line	End line	ANCOVA	ANCOVA
Treatment vs. Control	**0.336****	**0.331****	**0.333****	**0.358****	**0.301****	**0.373****	**0.291****	**0.393*****
	(0.134)	(0.146)	(0.132)	(0.153)	(0.118)	(0.142)	(0.119)	(0.138)
Baseline Score		-0.0240		-0.0597		0.0702		0.0541
		(0.132)		(0.142)		(0.113)		(0.115)
Treatment x Initial Cognitive Score (=1 if above median)[c]		0.0312		-0.00128		-0.208*		-0.275**
		(0.0932)		(0.0983)		(0.114)		(0.122)
Treatment x Initial Noncognitive Score (=1 if above median)[d]			0.0187	0.0200			0.0263	0.0556
			(0.0509)	(0.0516)			(0.0493)	(0.0513)
Constant	-0.171	-0.171	-0.138	-0.136	-0.153	-0.153	-0.135	-0.117
	(0.103)	(0.103)	(0.0968)	(0.0967)	(0.0911)	(0.0912)	(0.0990)	(0.100)
Observations	819	819	819	819	793	793	793	793
R-squared	0.028	0.028	0.032	0.033	0.023	0.028	0.024	0.032

Note: 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; ***, **, * denote at 1 percent, 5 percent, and 10 percent level respectively.
Note[a]: The level of agreement to the statement "I did well in the test", where responses were recorded on four-point scale:
Strongly Agree, Somewhat Agree, Somewhat Disagree, Strongly Disagree.
Note[b]: The level of agreement to the statement "I can confidently express my opinion", where responses were recorded on four-point scale:
Strongly Agree, Somewhat Agree, Somewhat Disagree, Strongly Disagree.
Note[c]: The Initial Cognitive Score stands for the DT Score for columns (6), (8)-(10) and PTSII Score for column (7).
Note[d]: The Initial Non Cognitive Score stands for the average non-cogntive score based on all 27 questions for columns (6) and (7). For columns (8)-(10), RSES, CPCS, and GRIT are used respectively.

Table 9: Impacts on KUMON Program on Students' Outcomes - Control for Longer Sessions

Impacts of KUMON Program on Students' Cognitive Outcomes Cross-sectional Heterogenous Specification & Longer Session														
	Grade 3					***Grade 4***					***All grades***			
Dependent Variables	DT_per_min [a]	PTS Score	RSES [b]	CPCS [b]	GRIT [b]	DT_per_min [a]	PTS Score	RSES [b]	CPCS [b]	GRIT [b]	PTS Score	RSES [b]	CPCS [b]	GRIT [b]
	(1)	(4)	(5)	(6)	(7)	(8)	(11)	(12)	(13)	(14)	(11)	(12)	(13)	(14)
Treatment vs. Control	**2.301*** **	**0.868** **	**0.0210**	**0.166**	**0.277***	**3.079** **	**0.990*** **	**0.305**	**0.409**	**0.0769**	**0.924*** **	**0.138**	**0.265**	**0.194***
	(0.445)	(0.339)	(0.224)	(0.198)	(0.141)	(1.200)	(0.297)	(0.291)	(0.292)	(0.157)	(0.231)	(0.179)	(0.168)	(0.104)
Treatment x Longer session	-0.301	-0.347	-0.113	-0.188	-0.404	-1.763	0.132	-0.209	-0.362	-0.442**	-0.199	-0.155	-0.253	-0.406*
	(0.549)	(0.450)	(0.325)	(0.306)	(0.309)	(1.222)	(0.285)	(0.333)	(0.308)	(0.156)	(0.323)	(0.244)	(0.231)	(0.219)
Constant	0.723***	1.146***	0.00852	-0.0561	-0.0670	0.979***	0.505***	-0.126	-0.139	0.0490	0.579***	-0.0590	-0.0896	-0.00874
	(0.138)	(0.164)	(0.119)	(0.130)	(0.116)	(0.245)	(0.107)	(0.119)	(0.101)	(0.0954)	(0.120)	(0.114)	(0.103)	(0.0859)
Number of Observations	473	501	496	496	493	338	336	336	336	324	837	832	832	817
R-squared	0.251	0.110	0.002	0.006	0.027	0.219	0.222	0.017	0.033	0.021	0.194	0.004	0.014	0.022

Note: 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; **, **, * denote at 1 percent, 5 percent, and 10 percent level respectively.

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

Note[b]: The Proficiency Test of Self Learning first half, which consists of 348 math questions. Among the 27 survey questions that Kumon has prepared, 10 are consistent with the Childrens Perceived Competence Scale (CPCS Non-Cognitive Score), 8 with the Rosenberg Self-Esteem Scale (RSES Non-Cognitive Score), and 3 with the Grit Scale (GRIT) (GRIT Non-Cognitive Score). For the survey questions related to each Non-Cognitive Score, see Appendix 1. The responses are recorded on 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 10: Impacts on KUMON Program on Students' Outcomes - Continuous Interaction terms

Impacts of KUMON Program on Students' Cognitive Outcomes Cross-sectional Heterogenous Specification & Continuous Variables

Dependent Variables	*Grade 3*					*Grade 4*					*All grades*			
	DT per min[a]	PTS Score	RSES [b]	CPCS [b]	GRIT [b]	DT per min [a]	PTS Score	RSES [b]	CPCS [b]	GRIT [b]	PTS Score	RSES [b]	CPCS [b]	GRIT [b]
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Treatment vs. Control	**2.141*****	**0.716****	**-0.0893**	**0.0322**	**0.0488**	**2.677*****	**1.181*****	**0.268**	**0.315**	**0.00220**	**0.882*****	**0.0686**	**0.163**	**0.0367**
	(0.322)	(0.281)	(0.173)	(0.156)	(0.134)	(0.810)	(0.193)	(0.220)	(0.220)	(0.164)	(0.194)	(0.137)	(0.130)	(0.0977)
Treatment x Initial Cognitive Score [c]	0.605*	0.345	0.158*	0.164*	0.124	0.241	0.575***	0.0814	0.0429	0.188	0.404*	0.124	0.118	0.154*
	(0.295)	(0.282)	(0.0881)	(0.0901)	(0.112)	(0.543)	(0.0967)	(0.173)	(0.174)	(0.126)	(0.212)	(0.0886)	(0.0904)	(0.0833)
Treatment x Initial Noncognitive Score [d]	0.172	-0.109	0.0834	0.0359	0.0556	0.475	0.0734	0.0308	0.0299	0.0638	-0.0440	0.0581	0.0298	0.0596*
	(0.160)	(0.157)	(0.0588)	(0.0383)	(0.0488)	(0.532)	(0.101)	(0.142)	(0.151)	(0.0517)	(0.105)	(0.0683)	(0.0661)	(0.0349)
Constant	0.811***	1.248***	0.127	0.0698	0.0892	1.015***	0.523***	-0.119	-0.134	0.0572	0.630***	-0.0173	-0.0456	0.0544
	(0.170)	(0.172)	(0.132)	(0.122)	(0.0877)	(0.252)	(0.110)	(0.121)	(0.103)	(0.0992)	(0.118)	(0.119)	(0.109)	(0.0895)
Number of Observations	473	501	496	496	493	338	336	336	336	324	837	832	832	817
R-squared	0.289	0.148	0.039	0.043	0.072	0.214	0.306	0.017	0.023	0.016	0.239	0.022	0.028	0.048

Note: 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; ***, **, * denote at 1 percent, 5 percent, and 10 percent level respectively.

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

Note[b]: The Proficiency Test of Self Learning first half, which consists of 348 math questions. Among the 27 survey questions that Kumon prepared, 10 are consistent with the Childrens Perceived Competence Scale (CPCS Non-Cognitive Score), 8 with the Rosenberg Self-Esteem Scale (RSES Non-Cognitive Score), and 3 with the Grit Scale (GRIT) (GRIT Non-Cognitive Score). For the survey questions related to each Non-Cognitive Score, see Appendix 1. 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.

Note[c]: The Initial Cognitive Score stands for the DT Score for columns (6), (8)-(10), and the PTSII Score for column (7).

Note[d]: The Initial Non Cognitive Score stands for the average non-cogntive score based on all 27 questions for columns (6) and (7). For columns (8)-(10), RSES, CPCS, and GRIT are used respectively.

Table 11: Impact of Kumon program on Student Cognitive Outcome: Difference-in-Differences Estimates

Impacts of KUMON Program on Students' Cognitive Outcomes DID Specification

	Grade 3				*Grade 4*				*All grades*	
Dependent Variables	Improve in DT per min [a] (1)	Improve in PTSII Score (2)	Improve in DT per min [a] (3)	Improve in PTSII Score (4)	Improve in DT per min [a] (5)	Improve in PTSII Score (6)	Improve in DT per min [a] (7)	Improve in PTSII Score (8)	Improve in PTSII Score (9)	Improve in PTSII Score (10)
Treatment vs. Control	**1.865*****	**0.877****	**1.522*****	**1.399*****	**2.733****	**1.584*****	**2.153*****	**1.775*****	**1.162*****	**1.572*****
	(0.295)	(0.397)	(0.352)	(0.381)	(0.975)	(0.324)	(0.719)	(0.313)	(0.276)	(0.264)
Treatment x Initial Cognitive Score (=1 if above median)[b]			0.381	-0.856**			-0.582	-0.661***		-0.814***
			(0.230)	(0.390)			(0.803)	(0.187)		(0.289)
Treatment x Initial Noncognitive Score (=1 if above median)[c]			0.346**	-0.197			1.669	0.127		-0.0926
			(0.145)	(0.295)			(1.093)	(0.142)		(0.204)
Constant	0.860***	1.059***	0.860***	1.059***	0.848***	0.226	0.848***	0.226	0.408*	0.385*
	(0.197)	(0.265)	(0.197)	(0.265)	(0.276)	(0.257)	(0.277)	(0.258)	(0.229)	(0.222)
Number of Observations	473	467	473	467	338	320	338	320	787	787
R-squared	0.181	0.110	0.190	0.169	0.205	0.322	0.240	0.345	0.224	0.267

Note: 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; ***, **, * denote at 1 percent, 5 percent, and 10 percent level respectively.

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

Note[b]: The Initial Cognitive Score stands for the DT Score for columns (6), (8)-(10), and the PTSII Score for column (7).

Note[c]: The Initial Non Cognitive Score stands for the average non-cogntive score based on all 27 questions for columns (6) and (7). For columns (8)-(10), RSES, CPCS, and GRIT are used respectively.

Table 12: Impact of Kumon program on Student Non Cognitive Outcome: Difference-in-Differences Estimates

Impacts of KUMON Program on Students' Non Cognitive Outcomes DID Specification

Dependent Variables	Grade 3						Grade 4						All grades					
	RSES [a]	CPCS [a]	GRIT [a]	RSES [a]	CPCS [a]	GRIT [a]	RSES [a]	CPCS [a]	GRIT [a]	RSES [a]	CPCS [a]	GRIT [a]	RSES [a]	CPCS [a]	GRIT [a]	RSES [a]	CPCS [a]	GRIT [a]
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
Treatment vs. Control	**-0.173**	**-0.383**	**-0.392***	**0.345**	**0.163**	**0.503***	**0.225**	**0.211**	**-0.336***	**1.069*****	**0.992*****	**0.0899**	**0.00935**	**-0.111**	**-0.367****	**0.672*****	**0.559*****	**0.301**
	(0.287)	(0.257)	(0.200)	(0.276)	(0.289)	(0.279)	(0.247)	(0.247)	(0.185)	(0.202)	(0.242)	(0.224)	(0.192)	(0.184)	(0.136)	(0.191)	(0.202)	(0.181)
Treatment x Initial Cognitive Score (=1 if above median)[b]				0.289	0.227	0.0796				-0.0171	-0.0487	0.385*				0.147	0.0821	0.208
				(0.181)	(0.192)	(0.220)				(0.208)	(0.223)	(0.190)				(0.144)	(0.153)	(0.155)
Treatment x Initial Noncognitive Score (=1 if above median)[c]				-1.314***	-1.180***	-1.599***				-1.561***	-1.403***	-1.305***				-1.398***	-1.275***	-1.470***
				(0.192)	(0.185)	(0.152)				(0.136)	(0.236)	(0.204)				(0.130)	(0.143)	(0.121)
Constant	0.232	0.365*	0.385***	0.232	0.365*	0.385***	-0.0758	-0.0407	0.216	-0.0758	-0.0407	0.216	0.0184	0.0998	0.229*	0.0317	0.0962	0.185
	(0.208)	(0.204)	(0.0974)	(0.208)	(0.204)	(0.0977)	(0.0676)	(0.0581)	(0.130)	(0.0678)	(0.0582)	(0.131)	(0.104)	(0.105)	(0.113)	(0.0933)	(0.0975)	(0.111)
Number of Observations	380	380	376	380	380	376	316	316	293	316	316	293	696	696	669	696	696	669
R-squared	0.004	0.018	0.023	0.131	0.126	0.250	0.007	0.006	0.015	0.158	0.127	0.116	0.001	0.003	0.020	0.134	0.114	0.185

Note: 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; ***, **, * denote at 1 percent, 5 percent, and 10 percent level respectively.
Note[a]: The Proficiency Test of Self Learning first half, which consists of 348 math questions. Among the 27 survey questions that Kumon prepared: 10 are consistent with the Childrens Perceived Competence Scale (CPCS Non-Cognitive Score), 8 with the Rosenberg Self-Esteem Scale (RSES Non-Cognitive Score), and 3 with the Grit Scale (GRIT Non-Cognitive Score). For the survey questions related to each Non Cognitive Score, see Appendix 1. 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.
Note[b]: The Initial Cognitive Score stands for the DT Score for columns (6), (8)-(10), and the PTSII Score for column (7).
Note[c]: The Initial Non Cognitive Score stands for the average non-cogntive score based on all 27 questions for columns (6) and (7). For columns (8)-(10), RSES, CPCS, and GRIT are used respectively.

Table 13: Teachers' Assessment Ability (1)

Correlation Values between Teachers' assessment and students' performance

Dependent Variables	All Sample	Grade 3 Students	Grade 4 Students
DT Score			
Teacher evaluation x (1-Treatment) x (1-End line)	0.397***	0.358***	0.391***
	(0.0579)	(0.0659)	(0.0950)
Teacher evaluation x Treatment x (1-End line)	0.193**	0.0721	0.454**
	(0.0830)	(0.0538)	(0.173)
Teacher evaluation x (1-Treatment) x End line	0.472***	0.553***	0.312***
	(0.0678)	(0.0802)	(0.0730)
Teacher evaluation x Treatment x End line	0.610**	0.546	1.382**
	(0.263)	(0.316)	(0.559)
Control Baseline = Treatment Baseline	**0.42**	**1.19**	**0.79**
Control End line = Treatment End line	**1.36**	**2.42**	**0.86**
Number of Observations	1,202	732	470
R-squared	0.520	0.577	0.515
PTSII Score			
TTeacher evaluation x (1-Treatment) x (1-End line)	0.124	-0.0358	0.369***
	(0.206)	(0.266)	(0.113)
Teacher evaluation x Treatment x (1-End line)	0.159	0.167	0.295***
	(0.118)	(0.130)	(0.0930)
Teacher evaluation x (1-Treatment) x End line	0.325***	0.332***	0.292***
	(0.0469)	(0.0665)	(0.0368)
Teacher evaluation x Treatment x End line	0.547***	0.543***	0.580***
	(0.0754)	(0.110)	(0.0734)
Control Baseline = Treatment Baseline	**0.03**	**0.17**	**0.11**
Control End line = Treatment End line	**7.68*** **	**5.04** **	**5.68** **
Number of Observations	1,284	762	522
R-squared	0.531	0.568	0.521

Table 14: Teachers' Assessment Ability (2)

Variance of Difference in Standardized Values

Dependent Variables	All Sample	Grade 3 Students	Grade 4 Students
DT Score			
Treatment	0.0301	0.162*	-0.122
	(0.0728)	(0.0898)	(0.0823)
End line	-0.0872	-0.168**	0.0896
	(0.0717)	(0.0745)	(0.121)
TreatmentxEnd line	**0.0186**	**-0.0118**	**-0.0391**
	(0.107)	(0.122)	(0.160)
Constant	1.132***	1.120***	1.146***
	(0.0494)	(0.0585)	(0.0844)
Number of Observations	1,202	732	470
R-squared	0.040	0.333	0.100
PTSII Score			
Treatment	0.0472	0.0895	-0.00807
	(0.0684)	(0.104)	(0.0838)
End line	-0.0489	-0.0217	-0.0873
	(0.0726)	(0.0788)	(0.144)
TreatmentxEnd line	**-0.114**	**-0.137**	**-0.0914**
	(0.0893)	(0.111)	(0.162)
Constant	1.095***	1.078***	1.118***
	(0.0545)	(0.0707)	(0.0876)
Number of Observations	1,284	762	522
R-squared	0.086	0.109	0.092

Note: 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;
***, **, * denote at 1 percent, 5 percent, and 10 percent level respectively.

6 Figures

Figure 1: Responses to all Survey Questions at Baseline

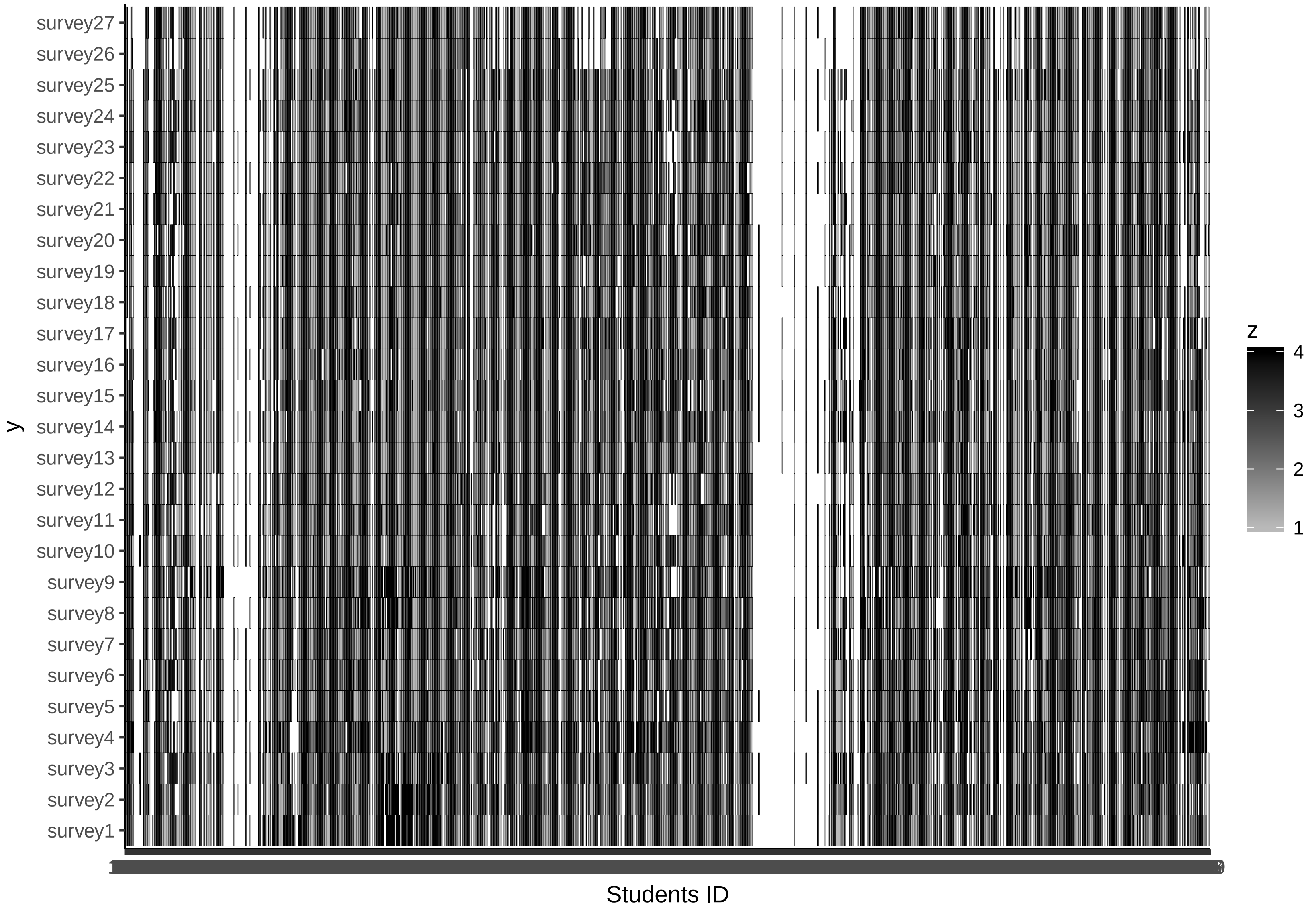
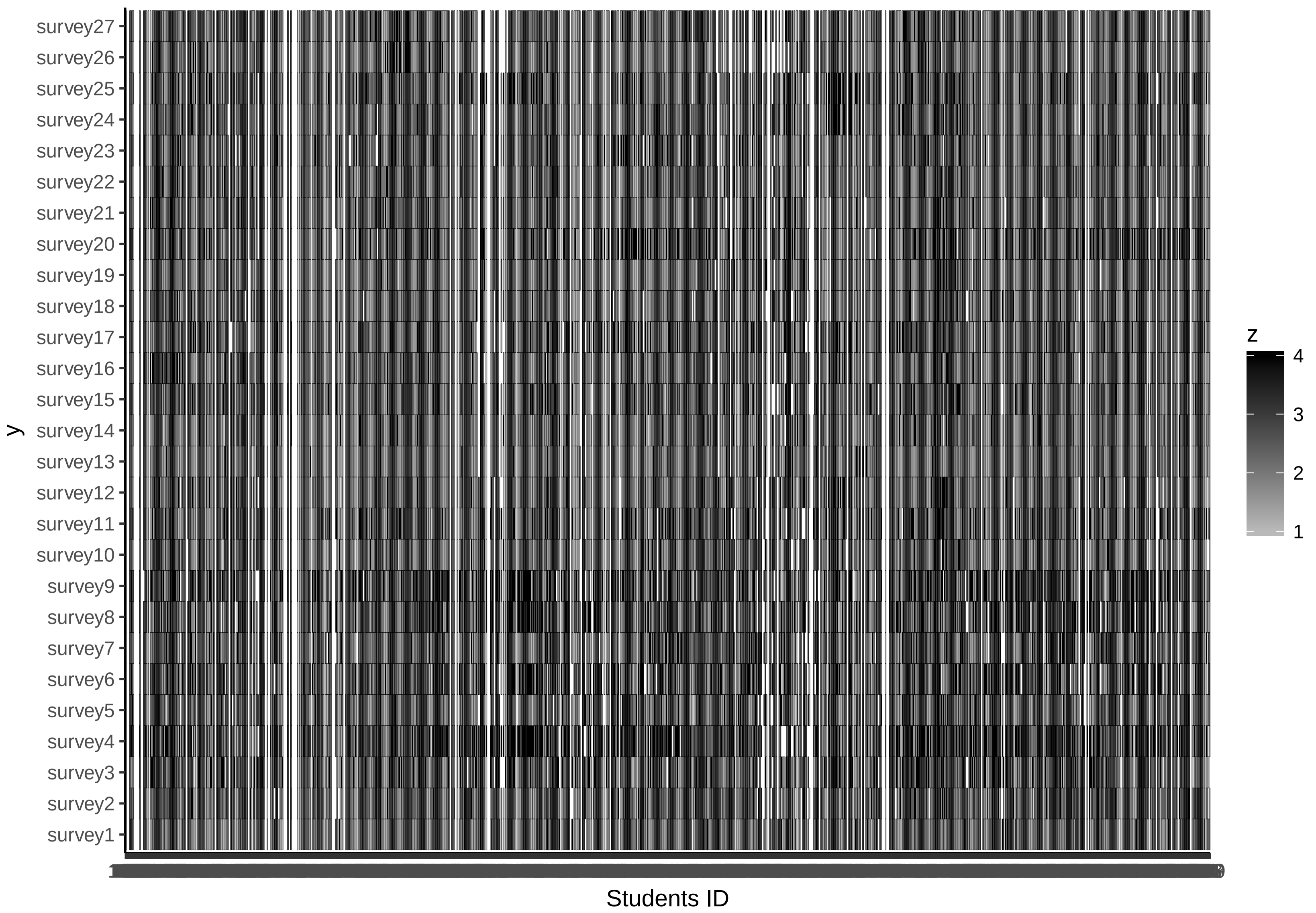


Figure 2: Responses to all Survey Questions at End line



Appendix 1: Non-Cognitive Ability Survey Questions

Table A1: PTS II second half of survey questions for measuring non-cognitive abilities

Number	Question in English	CPCS	RSES	GRIT
1	I did well in 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 dont think I am a helpful person.	x	x	
7	I can confidently express my opinion.	x		
8	I dont 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			
13	I like to do calculations.			x
14	I can calculate in my head when I go shopping.			x
15	I think speed is important when solving problems.			x
16	When studying, I believe everything will go well if I correctly follow instruction			
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 dont 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 dont go as expected.			
27	It doesnt matter whether I fail in the beginning because I believe that things will eventually work out.			

Note: Among the 27 survey questions that Kumon prepared, 10 are consistent with he Children's Perceived Competence Scale; CPCS (Sakurai (1992)Harter (1979)), 8 with the Rosenberg Self-Esteem Scale; RSES (Rosenberg (1965)), and 3 with the Grit Scale; GRIT (Duckworth). The rest are more specific to the Kumon-Method of learning original with 4 Bangladesh specific questions (question 24-27). The Japanese version of the original Kumon survey questions is based on Sakurai (1992).

Appendix 2: Data Cleaning and Merging

Sample Attrition: Table A2 shows that the baseline test scores are not correlated with the probability of being out of sample in the end line.

Table A2: 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	0.00122	0.00896	-0.00578	-0.0220
	(0.00498)	(0.0356)	(0.0117)	(0.0494)
Baseline PTSII Score	-0.00140	-0.0122	-0.00164	-0.00877
	(0.000974)	(0.00926)	(0.00222)	(0.0126)
Number of Observations	481	481	357	357
R-squared	0.008		0.017	

Cluster standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

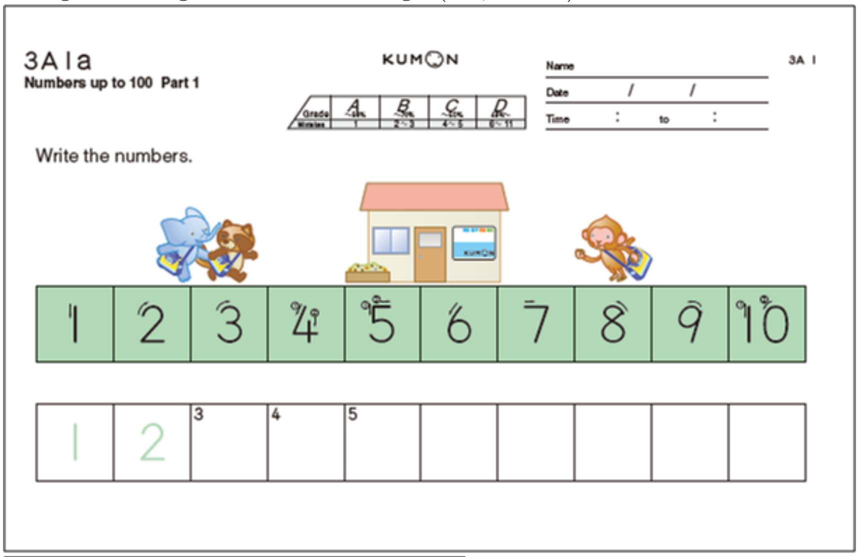
Data Merging: We used student number and school number which are uniquely assigned to each student and each school in our experiment to merge different datasets. Table A3 shows each data set and variables.

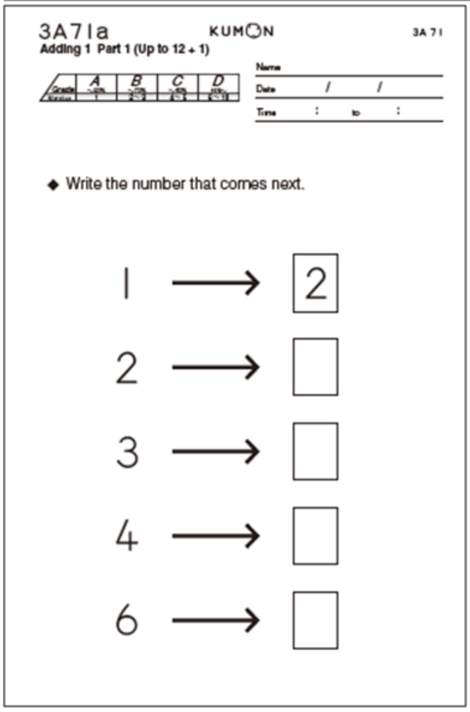
Table A3: List of Datasets

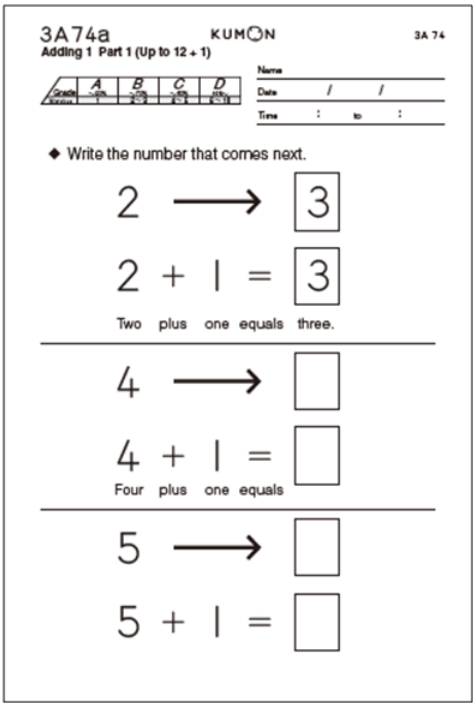
Dataset	Collected Time	Variables Used	Key Variables	Size
DT Score	Baseline	name_student, DOB, treatment, grade, name_branch, area, name_teacher, name_school, score_DTScoreP1, time_DTScoreP1, score_DTScoreP2, time_DTScoreP2, score_DTScoreP3, time_DTScoreP3	student_no, school_no	934
PTSII	Baseline	name_student, starting_level, starting_material, overall_score, survey1 survey27	student_no, school_no	905
Teacher Evaluation	Baseline	evaluation	student_no	934
DT Score	End line	name_student, DOB, treatment, grade, name_branch, area, name_teacher, name_school, score_DTScoreP1, time_DTScoreP1, score_DTScoreP2, time_DTScoreP2, score_DTScoreP3, time_DTScoreP3, student_gender	student_no, school_no	974
PTSII	End line	name_student, starting_level, starting_material, overall_score, survey1 survey27	student_no, school_no	837
Teacher Evaluation	End line	evaluation		738
Parents/Guardian Questionnaire		name_school, gender, age, relationship_hh_head, marital_status, educ_completed_grade, main_activity_past_month, hours_worked_day, days_worked_week, a11_last_income_drawn, frequently_income_drawn, grade_class, two_weeks_taught_discuss, books_suitable_child, often_read_story_child, agree_help_study, frequently_eat_meat_fish, frequently_eat_egg frequently_drink_milk, main_income_source, avgincome_range, electricity_connection, gas_connection, kind_toilet_facility, dwelling_type source_drinking_water, own_house, landamt_decimal	student_no	737

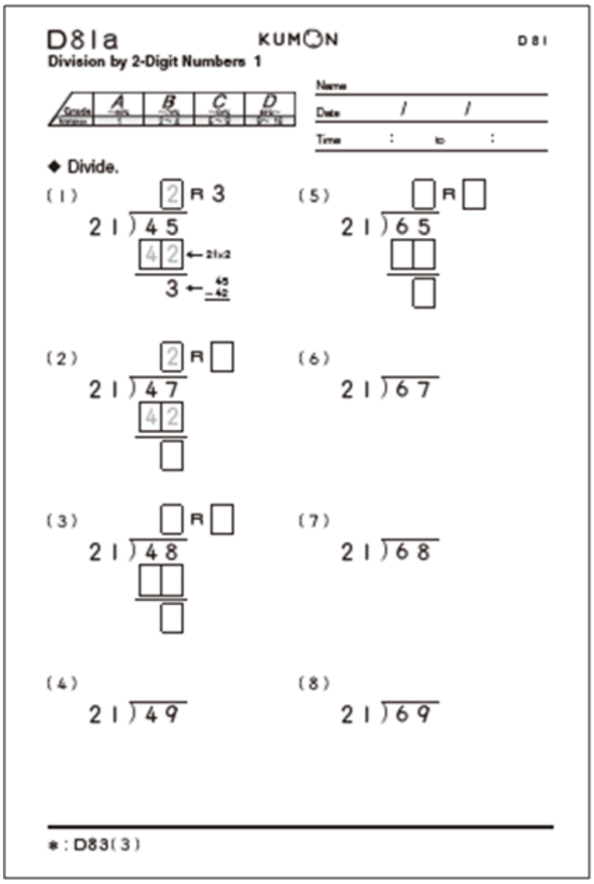
Appendix 3: Kumon Method Worksheet Examples

In the Kumon method, self-learning process is enforced by the examples and hints (first few questions with gray lines). Also, students only need to learn new math concepts and calculation steps in very small increments at each worksheet, which help them learn by themselves. For example, the first worksheet (3A1a) is letting students learn the order of numbers (up to 100 for example). Then after students have mastered these worksheets without an error within a targeted timeframe, they start to learn the concept of addition (note: a completion within a targeted time is a proxy for letting students advance to the next worksheet.). The second worksheet (3A71a) introduces students to a 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. Then finally, in the third worksheet (3A74a), students learn the concept of adding one using the summation sign (i.e., "+ 1").









The last worksheet (D81a) shows the division by 2-digit numbers. Even with more complicated arithmetic, the example and hits as well as the preceding worksheets make it possible for students to self-learn calculation skills and some of the math concepts behind it. Please note that these worksheets are the English versions. In the case of the BRAC primary school trail, all the materials are translated into Bengali, the local language that BRAC Primary School students regularly use in class.