Still Biased?

A Remaining Classical Selection Problem of RCTs in Education

By

Hikaru Kawarazaki (University College London)
Minhaj Mahmud (Asian Development Bank)
Yasuyuki Sawada (The University of Tokyo)
Mai Seki (Ritsumeikan University)
Kazuma Takakura (The University of Tokyo)

May 2022

CREPE DISCUSSION PAPER NO. 121

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

Still Biased?

A Remaining Classical Selection Problem of RCTs in Education

Hikaru Kawarazaki, Minhaj Mahmud, Yasuyuki Sawada, Mai Seki, and Kazuma Takakura*

6th May 2022

Abstract

In the already very rich and crowded literature on education interventions, the use of test scores to capture students' cognitive abilities has been the norm when measuring the impact. We show that even in randomized controlled trials (RCTs), estimated treatment effects on the true latent abilities can still be biased towards zero, because test scores are often censored outside of zero and full marks. This paper employs *sui generis* data from a field experiment in Bangladesh as well as data sets from existing highly-cited studies in developing countries to illustrate theoretically and empirically that this remaining classical sample selection problem exists. We suggest three concrete ways to correct such bias: First, to employ the conventional sample selection correction methods; second, to use tests that are designed with an extensive set of questions from easy to challenging levels which allow students to answer the maximum they could; and third, to incorporate each student's completion time in estimation.

JEL code: C24, I20, O15

*Hikaru Kawarazaki is a graduate student in Economics at University College London and a PhD scholar at the Institute for Fiscal Studies (hikaru.kawarazaki.20@ucl.ac.uk); Minhaj Mahmud is a senior economist at Asian Development Bank (mmahmud@adb.org); Yasuyuki Sawada is a professor of economics at the University of Tokyo (sawada@e.u-tokyo.ac.jp); Mai Seki is an associate professor in Economics at Ritsumeikan University (maiseki@fc.ritsumei.ac.jp); Kazuma Takakura is a graduate student at the University of Tokyo (takakura-kazuma190@g.ecc.u-tokyo.ac.jp). We thank Shotaro Beppu for his superb research assistance. Do not circulate without the authors' permission. We appreciate the comments from Costas Meghir and Imran Rasul.

1 Introduction

Randomized controlled trials (RCTs) in economics have been very popular in evaluating education interventions. The literature is already very rich and crowded in the context of both developed and developing countries (Kraft, 2020; Coe, 2002; Bloom et al., 2008; Lipsey et al., 2012; Kremer, Brannen and Glennerster, 2013; Ganimian and Murnane, 2016; Evans and Popova, 2015; McEwan, 2015; Glewwe, 2014; Duflo, Dupas and Kremer, 2011; Duflo, Hanna and Ryan, 2012; Duflo, Dupas and Kremer, 2015; Glewwe et al., 2004; Pradhan et al., 2014; Muralidharan, Singh and Ganimian, 2019). While many studies have investigated the effectiveness of different educational programs on improvements in children's cognitive ability captured by a variety of outcome measures, the most typical metric has been scores from standardized tests. Utilizing test scores is straightforward because it is believed to be a good measure of ability all over the world: PISA and TIMSS have been employed for international comparisons of cognitive abilities; and test scores are used for entrance exams in many countries (for example, A-level in the UK and SAT in the US) and also for proficiency tests in language and/or math (e.g., IELTS, TOEFL, GRE, and TOEIC). Although there is a general trend to incorporate measures other than test scores especially for school admissions such as an essay or a letter on personal backgrounds, the weights on cognitive test scores still seem very high.

Since test scores are limited inside a range between zero and full marks by nature and thus the true abilities are censored outside the range, even in RCTs estimated treatment effects of education interventions can still be biased towards zero. This paper shows theoretically and empirically that this remaining classical sample selection problem exists in the real world. While Angrist and Pischke (2008) states that *"the estimation of causal effects in experiments presents no special challenges whether y_i is binary, non-negative, or continuously distributed. The interpretation of the right-hand side changes for different sorts of dependent variables, but you do not need to do anything special to get the average causal effect,"* we believe that the issue we investigate is what Angrist and Pischke (2008) calls *"a rare case where the outcome variables is truly censored."*[1] Even among students with zero scores, some students have better abilities than others do. This is especially true if the test happens to be inappropriately difficult. In tackling the seemingly-rampant censoring problem in education, we employ data from a

[1] Angrist and Pischke (2008) points out that *"(d)o Tobit-type latent-variable models ever make sense? Yes, if the data you are working with are truly censored. True censoring means the latent variable has an empirical counterpart that is the outcome of primary interest".*

field experiment in Bangladesh as well as existing highly-cited studies from other developing countries.

We also need to understand the elements behind the existence of an upper bound and a lower bound of a score, i.e., full marks and a zero score, respectively. For example, GRE has an upper bound of 170 and a lower bound of 130 for Quantitative and Verbal components and the score range of TOEFL is from 0 to 120. There seem to be several legitimate reasons for this conventional test design. First, it is difficult to make a test without bounds by nature. To avoid someone obtaining a full score, examiners need to make just as many questions as no one can solve all of them. But this kind of test is simply very costly, time-consuming, and thus unrealistic. In some cases, examiners do not have original intentions to diversify examinees' scores. For example, a small quiz in the middle of a college course can offer full marks to all the students if a lecturer intends to make it sure that the quiz-takers understand the contents well. In this case, it does not make sense to prevent students from obtaining a full score. In addition, an examiner may have good reasons to provide difficult exam questions, making a large number of students receive null scores. In this case, the latency of the lower-bound test score would become salient.

If we use test scores with upper and lower bounds for a program evaluation, however, we may suffer from estimated causal effects biased towards zero. As many studies report positive effects, the bias will be downward. This bias arises from the classical sample selection problem, long discussed in the literature on the censored regression models (Amemiya, 1984; Greene, 2012). This paper offers both theoretical and empirical analyses on these issues in the context of developing countries, providing refined impact assessments on education interventions. To the best of our knowledge, the existing evaluation studies on education interventions, including those based on RCTs, have been silent about this potential bias. Hence, we believe our study makes a novel contribution to the already rich and crowded literature on the impact evaluation of education interventions.

We also discuss three concrete ways to correct such bias at least partly. First, we can employ the conventional sample selection correction method to mitigate this bias. The second one is to make an exam with easy but more than enough questions so that students will not obtain a zero score nor full marks. Although this is not entirely impossible as we will show in this paper, to design and conduct such an exam would be administratively very costly, The third way is to collect and use additional information of each student in an exam, i.e., solution "time." Most

exams set a time limit which seems to play a critical role. Some might argue that examinees try to maximize their test score given the time limit and therefore the time information has already been taken into account in the score. If this is the case, there is no benefit of incorporating time information in the analysis. However, if available, additional data on time to solve questions in an exam allows us to distinguish the following two types of examinees for example: Those who attain a certain score using the most of given time; and those who obtain the same score using a much shorter time. It is natural to think that these two types of examinees have different abilities with the latter dominating the former.[2][3] Therefore, incorporating time data have a potential to measure ability and estimate the treatment effects with better precision.

The rest of this paper is organized as follows. First, in Section 2, we construct a formal model to clarify the sample selection problem, followed by Section 3 on data and empirical implementation. The final section provides concluding remarks.

2 Theoretical Framework

2.1 Censoring of Test Score

We describe our theoretical framework of estimation bias arising from the upper and lower bounds of the test score. In this case, the treatment effects of a program that develops ability will systematically suffer from an estimation bias towards zero due to the classical sample selection problem.[4] For simplicity, let us consider the case in which the treatment effects are estimated by a standard RCT where the treatment and control groups are compared based only on endline data without loss of generality.

Let y_i^* be the true test score in the absence of the upper bound of the score. We postulate a linear model of generating the true test score as follows:

$$y_i^* = \alpha_0 + \alpha_1 a_i + u_i, \tag{1}$$

[2]May not all types of questions be adequate to focus on the time efficiency. For example, it may make more sense for task-type questions but not necessarily for questions requiring deep thinking.

[3]Although several papers measure labor productivity in terms of speed, few focus on speed in education. There is no doubt that thinking about a single topic deeply spending a lot of time develops people's cognitive abilities, but speed is also essential as we discussed above.

[4]Here, we assume there is a time limit. But this leads to the same discussion of the upper bound of the test score. However, having unlimited time limit for an exam is unrealistic and we propose a compromise which uses the (quasi) unlimited number of questions (consequently no upper bound of the test score) and a certain time limit.

where a_i is the unobserved ability of individual i and u_i is an independent error term. It is straightforward to presume that $\alpha_1 > 0$, showing the test score is a good proxy of the true ability. Furthermore, consider an intervention program that develops ability through the following function:

$$a_i = \gamma_0 + \gamma_1 d_i + e_i, \tag{2}$$

where d_i is a treatment status of individual i which takes one if i is treated and zero otherwise, and e_i is an error term. We take an example of an effective intervention i.e., $\gamma_1 > 0$, implemented by a randomized controlled trial so that $d_i \perp\!\!\!\perp u_i, e_i$. In this case, combining Equation (1) and (2), we have:

$$y_i^* = \beta_0 + \beta_1 d_i + \varepsilon_i, \tag{3}$$

where $\beta_0 := \alpha_0 + \alpha_1\gamma_0$, $\beta_1 := \alpha_1\gamma_1$, and $\varepsilon_i := \alpha_1 e_i + u_i$. Note that $d_i \perp\!\!\!\perp \varepsilon_i$ because $d_i \perp\!\!\!\perp u_i, e_i$. Also we have $\beta_1 > 0$ because $\alpha_1 > 0$ and $\gamma_1 > 0$. Here we assume that $\varepsilon_i \sim F$ with some cumulative distribution F with mean 0 and variance σ^2.In order to address the core problem, we introduce an upper and lower bound of the observed test score. Let y_i be the observed test score of individual i so that we have:

$$y_i = \begin{cases} \bar{y} & \text{if } y_i^* \geq \bar{y} \\ y_i^* & \text{if } y_i^* \in (\underline{y}, \bar{y}) \\ \underline{y} & \text{if } y_i^* \leq \underline{y} \end{cases} \tag{4}$$

where lower and upper bounds satisfy a condition, $\underline{y} < \bar{y}$. Considering the conditional expectation of y_i given d_i, we have:

$$\frac{\partial \mathbb{E}[y_i|d_i]}{\partial d_i} = \beta_1 Pr(y_i^* \in (\underline{y}, \bar{y})|d_i),$$

which is strictly smaller than the true parameter, β_1, if there are students who obtain a full score or a zero mark (Greene, 2012).[5] In this case, if we run a regression in a naive way using the following model:

$$y_i = \tilde{\beta}_0 + \tilde{\beta}_1 d_i + \tilde{\varepsilon}_i, \tag{5}$$

[5]For the derivation, see Theorem 19.4 of Greene (2012).

then we have:

$$\tilde{\beta}_1 = \frac{\partial \mathbb{E}[y_i|d_i]}{\partial d_i} \in (0, \beta_1)$$

with the assumption that $\beta_1 > 0$. Therefore, the OLS estimator will be biased downward in estimating the true treatment effects.

A straightforward method to correct this downward bias is to employ a Tobit model as Amemiya (1984) discussed. Assume that the error term ε follows the normal distribution independently with the mean 0 and the variance σ^2. Then, the conditional probability that y_i^* is in $(\underline{y}, t)$ with $t \leq \bar{y}$ is:

$$\begin{aligned} Pr(y_i^* \in (\underline{y}, t)|d_i) &= Pr\left(\frac{\underline{y} - \beta_0 - \beta_1 d_i}{\sigma} < \varepsilon_i < \frac{t - \beta_0 - \beta_1 d_i}{\sigma} \middle| d_i\right) \\ &= \Phi\left(\frac{t - \beta_0 - \beta_1 d_i}{\sigma}\right) - \Phi\left(\frac{\underline{y} - \beta_0 - \beta_1 d_i}{\sigma}\right), \end{aligned}$$

where Φ is a cumulative density function of the standard normal distribution. Taking the derivative of the above conditional probability with respect to t and evaluating this at $t = y_i$, we have:

$$\frac{1}{\sigma}\phi\left(\frac{y_i - \beta_0 - \beta_1 d_i}{\sigma}\right).$$

where ϕ is a density function of the standard normal distribution. Since the probabilities that $y_i = \underline{y}$ and that $y_i = \bar{y}$ are $\Phi\left(\frac{\underline{y} - \beta_0 - \beta_1 d_i}{\sigma}\right)$ and $1 - \Phi\left(\frac{\bar{y} - \beta_0 - \beta_1 d_i}{\sigma}\right)$, respectively, the log-likelihood function can be written as follows:

$$\begin{aligned} \mathcal{L}(\beta_0, \beta_1; y_i, d_i) = \prod_{i=1}^{n} \Bigg[&\left\{\Phi\left(\frac{\underline{y} - \beta_0 - \beta_1 d_i}{\sigma}\right)\right\}^{\mathbb{1}\{y_i = \underline{y}\}} \left\{\frac{1}{\sigma}\phi\left(\frac{y_i - \beta_0 - \beta_1 d_i}{\sigma}\right)\right\}^{\mathbb{1}\{y_i \in (\underline{y}, \bar{y})\}} \\ &\times \left\{1 - \Phi\left(\frac{\bar{y} - \beta_0 - \beta_1 d_i}{\sigma}\right)\right\}^{\mathbb{1}\{y_i = \bar{y}\}} \Bigg], \end{aligned} \tag{6}$$

where $\mathbb{1}$ is an indicator function which takes one if the argument is true and zero otherwise. The parameters, β_0 and β_1, can be estimated by the maximum likelihood method.

Another method is using the least absolute deviations (LAD) estimation. Powell (1984) develops a technique to apply the LAD to a censored regression. One of the features of LAD is using the median, which allows weaker assumptions on the error term than in Tobit regression.

The model is given as follows:

$$Y_i^* = \boldsymbol{X}_i'\boldsymbol{\beta} + u_i \tag{7}$$

$$Y_i = \max\{Y_i^*, 0\} \tag{8}$$

$$\text{Med}[u_i|\boldsymbol{X}_i] = 0, \tag{9}$$

where the observation is censored at the lower bound, 0. In this case, we do not have to assume the distribution form of the error term u_i. The censored LAD estimator is given as follows:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{b}} \sum_{i=1}^{N} |Y_i - \max\{Y_i^*, 0\}|. \tag{10}$$

We also employ this LAD estimation method in the empirical analysis below so that we can compare the results of existing studies based on the OLS, the Tobit model, and the LAD estimation model.

2.2 Time Use

A child's true ability can be observed not only by his/her test score but also by time to solve an exam. Let p_i and t_i be, respectively, the test score in points and the speed, i.e., the time spent by each individual i to solve one question. Denote that $T > 0$ is the time limit of the exam. Let y_i be the total score, a proxy of ability based on both the score point and time information. To formally present these considerations, we postulate the following "true" total test score function in which y_i is a function of p_i and t_i:

$$y_i^* = s(p_i, t_i). \tag{11}$$

Note that a usual "observed" test score measure can be expressed as $\tilde{y}_i = s(p_i, T)$ given that all the examinees are expected to stay in the exam until the very end of the exam. In this setting, if all examinees can solve all the questions within the exam time, the speed does not directly affect the observed outcome. It would be natural to assume $\frac{\partial s}{\partial p_i} > 0$ and $\frac{\partial s}{\partial t_i} < 0$, because if you obtain a higher test score and spend less time than others in solving questions, then your ability should be regarded as better and so should the final total score be. Assume that the marginal rate of substitution, defined by $-\frac{\partial s/\partial p_i}{\partial s/\partial t_i}$, is positive and increasing. This makes the

shape of indifference curve, which shows the same ability line, as indicated in Figure 1.[6] This restriction incorporates the standard properties of the microeconomic model with two goods in which one component is a "bad" because more time indicates worse ability. Hence, our setting would be seen as natural. Since test score points and time to solve questions are determined by each student's (cognitive) ability, we can postulate that p_i and t_i are a function of ability, a_i, we have:

$$p_i = p(a_i)$$
$$t_i = t(a_i),$$

where we can assume:

$$\frac{\partial p(a_i)}{\partial a_i} > 0$$
$$\frac{\partial t(a_i)}{\partial a_i} < 0.$$

As mentioned above, the last two assumptions seem reasonable: If your ability is higher, you will obtain a higher score, which corresponds to $\alpha_1 > 0$ in Equation (1). Also, you can solve questions faster in this case.

Finally, we incorporate a treatment status of a program, d_i, which is assumed to be continuous without loss of generality. For example, we can think about a situation where we model the impact of continuous treatment intensity.[7] Assume that there is a positive treatment on the ability as in Equation (2) with $\gamma_1 > 0$: and $\frac{\partial a_i}{\partial d_i} > 0$.

In sum, we can represent the (continuous) treatment effects of the program on the total

[6]This figure is created for measuring students' ability by a test called Diagnosis Test when they start a program offered by Kumon Ltd., one of the most extensive non-formal education program founded in Japan and running their business across 59 countries including the UK as of November 2021, according to the Kumon Institute of Education Co., Ltd. See `https://www.kumongroup.com/eng/about/?ID=eng_about-kumon` for the details (Last access: 21st January, 2022). See the discussion in Section 3 for the details.

[7]If we consider a typical program evaluation setting, alternatively we can model the binary treatment by modifying this formula in a discrete way. We can then compare the difference in the derivatives of the total score with respect to the ability evaluated by each treatment status.

score for y_i and $\tilde{y}_i$ as follows:

$$\frac{\partial y_i}{\partial d_i} = \underbrace{\frac{\partial s_i}{\partial p_i}}_{(+)}\underbrace{\frac{\partial p_i}{\partial a_i}}_{(+)}\underbrace{\frac{\partial a_i}{\partial d_i}}_{(+)} + \underbrace{\frac{\partial s_i}{\partial t_i}}_{(-)}\underbrace{\frac{\partial t_i}{\partial a_i}}_{(-)}\underbrace{\frac{\partial a_i}{\partial d_i}}_{(+)}$$

$$\frac{\partial \tilde{y}_i}{\partial d_i} = \frac{\partial s_i}{\partial p_i}\frac{\partial p_i}{\partial a_i}\frac{\partial a_i}{\partial d_i}.$$

Therefore,

$$\frac{\partial y_i}{\partial d_i} > \frac{\partial \tilde{y}_i}{\partial d_i}. \tag{12}$$

Hence, if we measure the treatment effects only with the conventional test score without consideration of time information, we would underestimate the true treatment effect especially when the duration of the exam is too short and/or the number of exam questions are too few for which $t_i = T$ for all i. This model illustrates that measuring both score and time seems indispensable to accurately capture the treatment effects on educational outcomes.[8] We examine this theoretical implication empirically based on observed test scores and time spent from Sawada et al. (2020) in Section 3.

3 Data and Empirical Analysis

To show the extent to which the bias discussed above appears and how we can correct it at least partly, we employ two different sets of data sources: First, data from the highly cited empirical studies with randomized controlled trials (RCTs); and, second, the data set we collected for another study.

First, after replicating the three highly-impacted papers, i.e., Duflo, Dupas and Kremer (2011); Duflo, Hanna and Ryan (2012); Duflo, Dupas and Kremer (2015); Pradhan et al. (2014), we correct their estimators based on the above-mentioned existing methods of the Tobit and LAD models. Second, we show the effectiveness of the other approaches using the extended dataset with individual exam completion time taken from Sawada et al. (2020).

[8] One crucial note is that by thinking about a partial derivative, we assume that examinees will not change their behaviors to obtain better scores between two cases. If they knew the final score might increase when they sacrifice the quality of the answer but solve questions in a short time, the treatment effects in terms of test score might be underestimated. Therefore, to implement fully, it is essential to introduce correct students' incentives not for them to do so. However, we can also consider this type of strategies as an indication of a "better" ability. This relates to the discussion on how to utilize the information on time into the final measure of abilities, such as in Figure 1.

3.1 Replication and Correction of the Three Existing Studies

First, we replicate Duflo, Dupas and Kremer (2011) which examines the effects of the tracking system based on an RCT conducted with primary schools in Kenya. The study finds that such a system can generate positive impacts on students' test scores directly and indirectly. Figure 2 shows histograms of the standardized total score, math score, and literature score based on the data set of Duflo, Dupas and Kremer (2011). All of these figures show a mass at the lower bound of the test score, which would potentially generate the sample selection bias. Table 1 shows the estimated treatment effects of their program with the measure of the total score, math score, and literature score, and compares the estimated effects using the Ordinary Least Squares (OLS), which is the replication of their original results, and type I Tobit model, which considers and corrects the selection bias. We can verify that the OLS estimates are systematically lower than the Tobit estimates in most cases, especially for literature scores for which clustering at the lower bound test score is salient (Figure 2).[9][10]

Second, we reproduce the reported results in Duflo, Hanna and Ryan (2012) on India which investigates whether teachers reduce their absence if they have an incentive to work, which will result in the improvement of education standards. They conducted an RCT to examine the effects of financial incentives for teachers on students' test scores in elementary schools, using a linear regression model. As indicated in Figure 3, we can see that the distribution of the math score, retrieved from the original data, has a mass on the lower bound of the test score. This might be a potential source of downward bias in the absolute value of the estimated treatment coefficient. Table 2 shows the replicated results and our results based on the Tobit model. Especially, the estimated treatment effect of 0.271 with selection correction in Column (2) is 29% larger than the replicated coefficient of 0.210 in Column (1). A similar tendency appears in the language and total scores. Given that the estimated treatment effects in education interventions are generally small, around 0.2, this difference arising from selection bias seems substantial which needs some care. Column (3) reports the results based on the LAD estimation method.[11] It would be straightforward to see that the estimated impact is corrected

[9] We are now updating our manuscript by conducting another method called LAD, described in Section 2.1, which does not require the normality assumption.

[10] For further investigation, we are now conducting several tests on whether the original estimates and the correct estimates are statistically different or not. This information would be helpful to understand further on the potential bias.

[11] Note that the standard errors of the LAD are not clustered, so in this draft, we will not compare the significance level between ones from the LAD and others, while ones from the OLS and the Tobit model are clustered and we can compare them.

upwards by the LAD model.[12]

Third, we replicated the results of Duflo, Dupas and Kremer (2015), which examines the effects of the Extra Teacher Program (ETP) and the School-Based Management (SBM) on students' test scores, school attendance, and dropout. The paper employs three test score measures, i.e., total score, math score, and literacy score, in which we can see that the math score has a mass on the lower bound of the score as shown in Figure 4. Table 3 shows the replicated results and our results with selection correction. According to this table, the effects of the basic ETP program and the ETP with the SBM program increase by 6.8% and 4.7%, respectively, after selection correction. While the difference is modest, we do observe the selection bias.[13]

Finally, we re-estimated the same model with the same data from Pradhan et al. (2014), The paper examines the effects of strengthening school committees in public schools in Indonesia, based on an RCT. The histogram of the outcome variable, the standardized test score, is shown in Figure 5. Based on these histograms, we do not see obvious bunching especially in lower bounds. Therefore, if we use the correction method, we would expect there would be no change in the estimated treatment effects. This exercises could be taken as a sort of falsification test. Indeed, as we can see in Table 4, we do not see an apparent change in the original findings. This would suggest that it will be useful to conduct the Tobit model estimation for an RCT-based analysis regardless of the obvious existence of bunching in lower or upper bounds of the outcome variable..

3.2 Using Unique Test Features of Sawada et al. (2020)

Furthermore, to illustrate other ways to correct the potential bias, we employ test results collected by an RCT in Bangladesh as described in (Sawada et al., 2020). This RCT was conducted to evaluate the effectiveness of an education program, an individualized self-learning program, of Kumon Ltd. (hereafter Kumon), a world-widely famous non-formal education firm, among the disadvantaged students studying at BRAC Primary School (BPS). Our original subjects are around 1,000 students in 34 schools (i.e., about 30 students per school) which are

[12]The LAD gives smaller estimates for outcomes with a smaller mass at the point censored, so it would be important to see the distribution of the outcome of interest before naively running the OLS, in addition to comparing results from several methods. The same is true for some re-estimates of Duflo, Dupas and Kremer (2015) and Pradhan et al. (2014). Again, this seems to be because of the original empirical distribution of the test scores. What is important is to choose appropriate methods based on the shape of the distribution and carefully compare the results from several methods.

[13]Some results show smaller estimates for the Tobit model. This could be due to the normality assumption, which might not be plausible in this setting. However, we believe that it is important to compare the results using this method to quantify the potential significance of the bias.

randomly selected out of 179 schools in 4 branches of BRAC in the Dhaka area with grade-specific strata. Out of 34 schools, we randomly set 17 treated schools and 17 control schools (Sawada et al., 2020).

As a baseline data set, Kumon conducts Diagnosis Test (DT) for their individualized self-learning program to adjust the initial level for each student. DT score plays a critical role in the program because, to offer the learning materials at a suitable level for each child, it is needed to accurately determine the initial level of students' abilities. By offering DT, the program utilizes not only the mere test score but also the time spent to solve the problem set to evaluate students' abilities.[14] The full score of the DT was set at 70. The histogram of the DT score is shown in Figure 6. As we can observe, the data has an upper bound score where many observations are clustered.

In addition to the DT test score, Sawada et al. (2020) collected data on another test, called the Proficiency Test of Self Learning (PTSII) from which we only use the cognitive test part of PTSII (hereafter, PTSII-C) for our analysis. This test has a very large number of questions, 228 maths questions, with the time limit of 10 minutes. Therefore, by construction, no one is expected to finish solving all the questions in practice, meaning that there is no upper bound of the observed test score. At the same time, some of the questions are extremely easy, such as just tracing a line with a pencil, so essentially no one will obtain zero scores among primary school students.[15] As we expect, the histogram of PTSII-C test scores does not have lower and upper bounds (Figure 7). This corresponds to the setting discussed in Section 2.1.[16]

3.2.1 Sample Selection Bias of RCTs

Based on the theoretical framework discussed in Section 2.1, we provide evidence on how different the treatment effects could be if we have a bound of the test score. specifically, we run the following regression model:

$$w_i = \beta_0 + \beta_1 d_i + \varepsilon_i,$$

where the observed test score, w_i equals to y_i^* if we use PTSII-C because it does not have lower and upper bounds by construction (Figure 7) and $w_i = y_i$ if we use DT, because it has an upper bound (Figure 6).

Our strategy is to compare several measures of abilities and discuss how we can utilize the

[14] See Sawada et al. (2020) for the detail of the exam and program.
[15] Indeed, we do not observe anyone who obtained a zero score in our sample.
[16] Test A corresponds to DT and test B to PTSII-C.

information on test scores and time. Table 5 shows the summary statistic of the DT score per minute, DT score, DT time, PTSII-C score, and PSC math test scores. All of these variables (except for the PSC, which were taken at the endline) show a balance of the baseline average values which basically validate our RCT implementation.[17]

In Table 6, we report the estimated treatment effects of the Kumon program on cognitive abilities, measured by PTSII-C, DT score, DT time, and DT score per minute. As discussed above and shown in Figure 7, PTSII-C does not have a lower or upper bound of the test score. In this case, the estimated treatment effects by the OLS would capture the unbiased estimate. On the other hand, the OLS result using DT score, which has an obvious upper bound in Figure 6, would generate downward-biased treatment effect. Indeed, The reported DT Score and PTSII-C results in Table 6 are consistent with the existence of sample selection bias of the estimate based on the DT test scores.[18]

One way to mitigate sample selection bias is to adopt the classical Tobit model with a set of normality assumptions. Table 7 shows the comparison of the estimates of treatment effects measured by DT score using OLS and type I Tobit model. Here, out of 743 observations, 57 reached the upper bound. We can see the estimate of OLS is smaller in an absolute sense than that of the Tobit model.[19] This result also indicates possible downward bias in estimating the true treatment effect.

3.2.2 Top-Coding and True Censoring

Angrist and Pischke (2008) lists the American Current Population Survey (CPS) earnings data, which *top-codes* (censors) very high values of earnings to protect respondent confidentiality, as a leading example of *true censoring*, in which the latent variable has an empirical counterpart that is the outcome of primary interest. To text the validity of our approach to a top-coded data set, we perform an additional analysis with PTSII-C which does not have a mass at the bounds (Figure 7). We artificially top-code PTSII-C data by setting an upper bound test score at two alternative values, 60 or 80 in the raw score (Figure 8). Theoretically speaking, these artificially top-coded data would generate downward bias in the estimated treatment effect which can be

[17] See Sawada et al. (2020) for the further discussions. Table J4 is the corresponding table.

[18] Although we discuss the importance of incorporating the time information in Subsection 3.2.3, it is worth mentioning it here. As shown in Figure 9, the DT score per minute, which is the DT score divided by DT time, does not suffer from censoring. Taking several dimensions into account might mitigate this type of censoring issue so that the OLS estimator can be unbiased.

[19] Although the difference seems small, the effect is shown in the unit of standard deviation and an educational context, a 0.04 s.d. change is large. Therefore, this downward bias is crucial.

corrected if we employ the Tobit model.

Our actual data and analysis support these prior predictions. According to the estimation results reported in Panel A of Table 9, we can verify the downward biases in estimated treatment effects with top-coded data (Columns (2) and (3)). Also, as we expect, the magnitude of the bias is larger with a wider range of top-coding (Column (3)) is larger than that with a narrower range of top-coding (Column (3)). In testing the difference between these estimated coefficients, we reject the null hypothesis of the equality of these coefficients (columns (2) and (3) in panel C. These results indicate that the bias arising from censoring of test scores is not negligible.

We adopt the Tobit model to correct the bias whose estimation results are reported in Columns (2) and (3) in Panel A. We can observe that the estimation biases are mitigated. Indeed, formal tests of the null hypothesis in which the original OLS estimate with non top-coded data is equal to the Tobit estimate with top-coded data cannot be rejected statistically (Panel C). These results suggest that the Tobit model can effectively mitigate selection bias due to censoring.

3.2.3 Time Use

Based on the model formulated in Section 2.2, we present an analysis incorporating the time to solve the exam questions.

First, in Table 6, it is notable that Kumon mehotd reduced the solution time of the DT test significantly. Hence, incorporating solution time into the estimation of treatment effects on test scores can potentially mitigate selection bias. Indeed, the DT score per minute also shows statistically significant treatment effects. Also, as illustrated in Figure 9, it does not seem to suffer from censoring. Therefore, using the time information could have the potential to solve the issue of bias.

We can validate our framework by using the nation-wide Primary School Certificate (PSC) exam as our benchmark outcome. Table 8 shows the result of empirical analyses where we regress the GPA of PSC on DT score, DT time, and DT scores per minute at the beginning of the intervention.[20] Note that the last measure, DT score per minute, incorporates both time and score information. This table illustrates how precisely each variable can predict the later educational outcome, measured outside of the Kumon program, i.e., PSC exam GPA.[21] Note

[20]Note that we did not use the clustering here because our purpose is to predict the PSC based on the baseline outcomes.

[21]Note that PSC itself is measured by the score and not based on time. This approximation is a potential

that all variables are standardized. The reported coefficients indicate how well the prediction of PSC results based on each variable. Comparing Columns (1) and (3), which correspond to the predictions based only on the score and on both score and time, respectively, we can see that the latter can provide a more precise prediction because the coefficient is closer to unity with better fitness than the former.[22]

4 Concluding Remarks

In this paper, we show that even in randomized controlled trials (RCTs), estimated treatment effects can still be downward-biased due to the classical sample selection problem arising from the existence of the upper and lower bound of the observed test scores. We provide theoretical backgrounds of the mechanism of the underestimation. We then propose possible ways to mitigate such a bias. In addition to the classical sample selection correction method using the Tobit model, we can also incorporate information about the speed of solving questions. Even if it is difficult to measure the time to solve the questions due to the standard setting of the exams, usage of tests without an upper and lower bound of the test score would be an effective alternative to measure the abilities accurately.

In sum, we suggest three concrete ways to correct such bias. First, by employing the conventional sample selection correction methods; second, by using tests that are designed with an extensive set of questions from easy to challenging levels which allow students to answer the maximum they could; and third, by incorporating each student's completion time in estimations. Empirical results based on our experiment as well as an influential RCT confirm the existence of sample selection bias. Also the proposed adjustment methods can mitigate this classical selection bias at least partly. Especially, We conclude that using the time information when we measure the ability seems very effective.

Our theoretical and empirical results provide important implications for future studies. Although measuring people's cognitive abilities by test score is very common in a wide variety of existing studies in program evaluations and elsewhere, we need to be cautious in interpreting the estimation results because the estimated treatment effects might involve downward bias.

caveat, but this is the only measure available outside of the Kumon program.

[22]The p-value of the test which examines whether the two coefficients are the same or not is 0.0164.

References

Amemiya, Takeshi. 1984. "Tobit Models: A Survey." *Journal of Econometrics*, 24(1-2): 3–61.

Angrist, Joshua D, and Jörn-Steffen Pischke. 2008. "Mostly harmless econometrics." In *Mostly Harmless Econometrics*. Princeton university press.

Bloom, Howard S, Carolyn J Hill, Alison Rebeck Black, and Mark W Lipsey. 2008. "Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions." *Journal of Research on Educational Effectiveness*, 1(4): 289–328.

Coe, Robert. 2002. "It's the effect size, stupid." Vol. 12, 14.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." *American Economic Review*, 101(5): 1739–74.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2015. "School Governance, Teacher Incentives, and Pupil–Teacher Ratios: Experimental Evidence from Kenyan Primary Schools." *Journal of Public Economics*, 123: 92–110.

Duflo, Esther, Rema Hanna, and Stephen P. Ryan. 2012. "Incentives Work: Getting Teachers to Come to School." *American Economic Review*, 102(4): 1241–78.

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

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

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

Glewwe, Paul, Michael Kremer, Sylvie Moulin, and Eric Zitzewitz. 2004. "Retrospective vs. Prospective Analyses of School Inputs: the Case of Flip Charts in Kenya." *Journal of Development Economics*, 74(1): 251–268.

Greene, William H. 2012. *Econometric Analysis.* Pearson Education Limited.

Kraft, Matthew A. 2020. "Interpreting effect sizes of education interventions." *Educational Researcher*, 49(4): 241–253.

Kremer, Michael, Conner Brannen, and Rachel Glennerster. 2013. "The Challenge of Education and Learning in the Developing World." *Science*, 340(6130): 297–300.

Lipsey, Mark W, Kelly Puzio, Cathy Yun, Michael A Hebert, Kasia Steinka-Fry, Mikel W Cole, Megan Roberts, Karen S Anthony, and Matthew D Busick. 2012. "Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms." *National Center for Special Education Research.*

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

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

Powell, James L. 1984. "Least Absolute Deviations Estimation for the Censored Regression Model." *Journal of Econometrics*, 25(3): 303–325.

Pradhan, Menno, Daniel Suryadarma, Amanda Beatty, Maisy Wong, Arya Gaduh, Armida Alisjahbana, and Rima Prama Artha. 2014. "Improving Educational Quality through Enhancing Community Participation: Results from a Randomized Field Experiment in Indonesia." *American Economic Journal: Applied Economics*, 6(2): 105–26.

Sawada, Yasuyuki, Minhaj Mahmud, Mai Seki, An Le, and Hikaru Kawarazaki. 2020. "Fighting the Learning Crisis in Developing Countries: A Randomized Experiment of Self-Learning at the Right Level." CIRJE Discussion Papers CIRJE-F-1127, University of Tokyo.

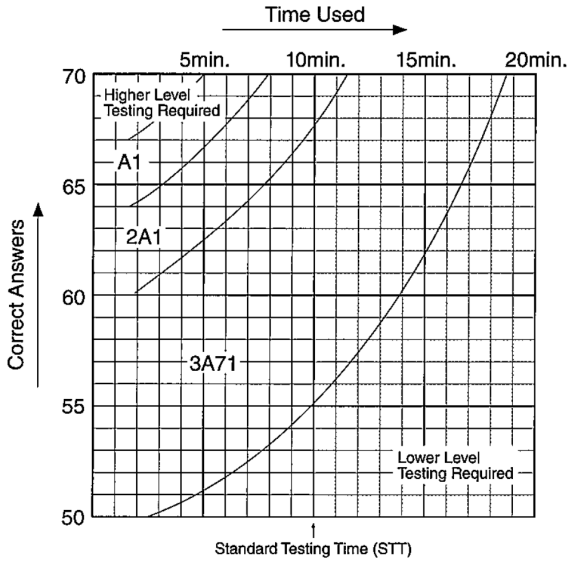


Figure 1. Diagnostic Test from Kumon Ltd.

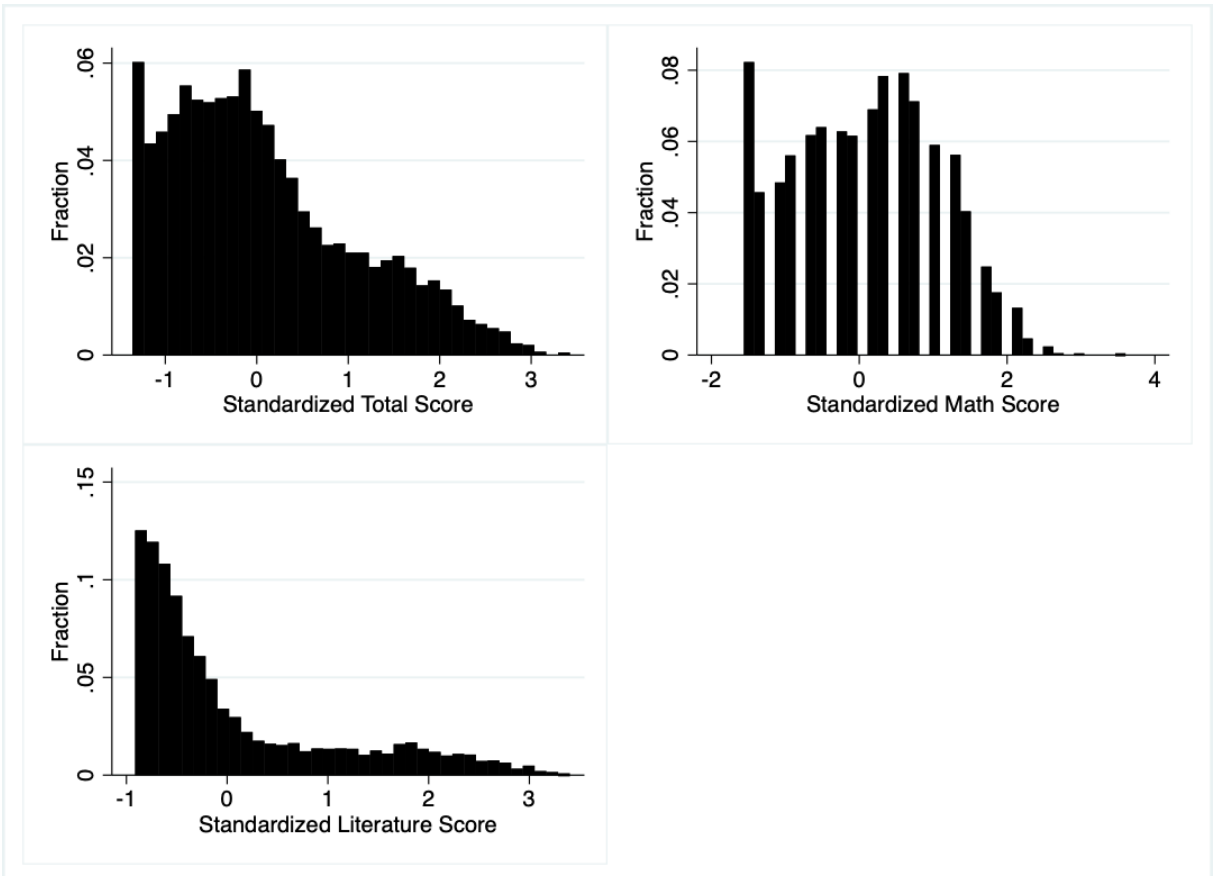


Figure 2. Distribution of Scores in Duflo, Dupas and Kremer (2011)

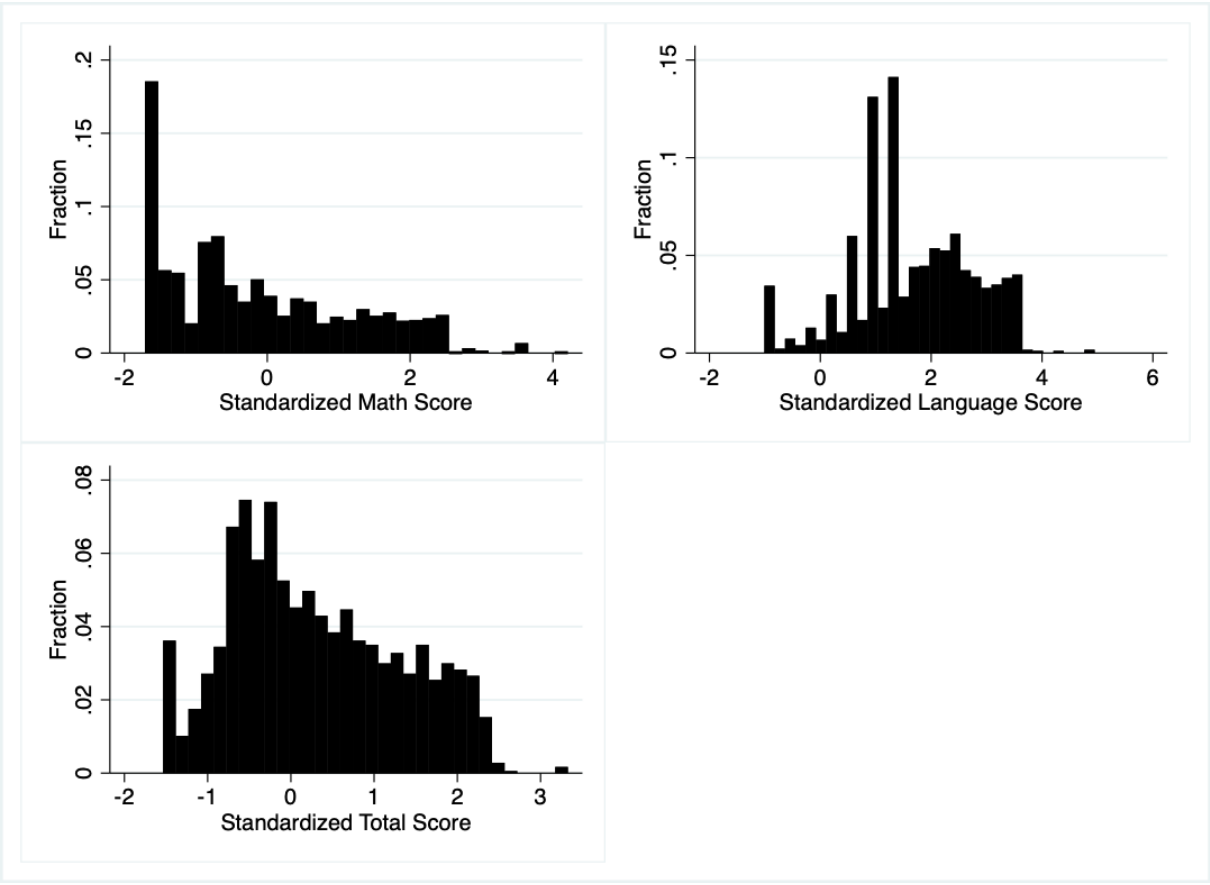


Figure 3. Distribution of Scores in Duflo, Hanna and Ryan (2012)

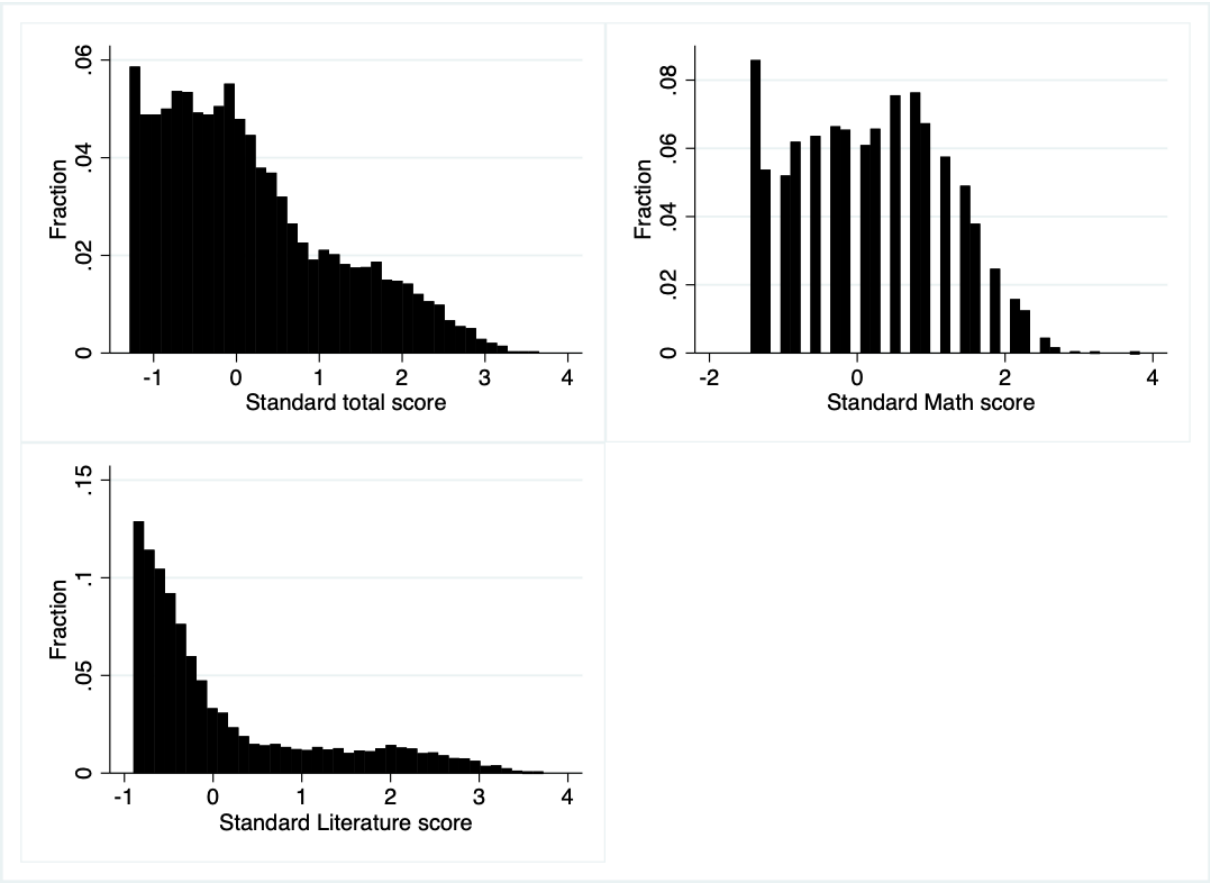


Figure 4. Distribution of Scores in Duflo, Dupas and Kremer (2015)

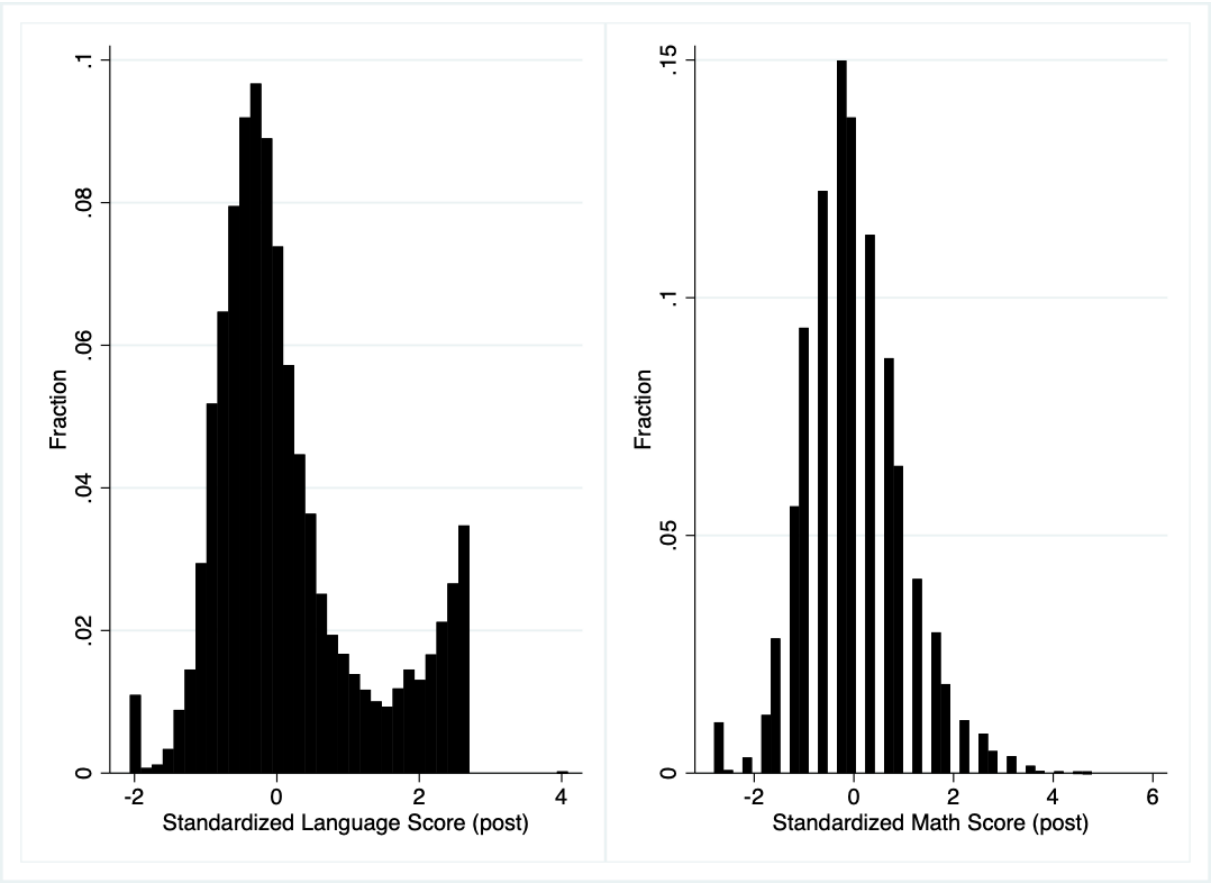


Figure 5. Distribution of Scores in Pradhan et al. (2014)

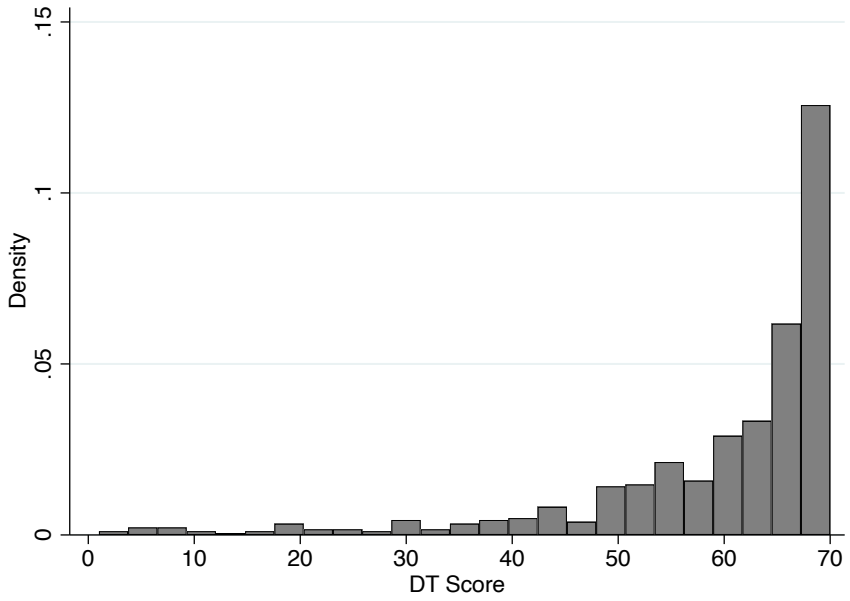


Figure 6. Distribution of DT score (endline)

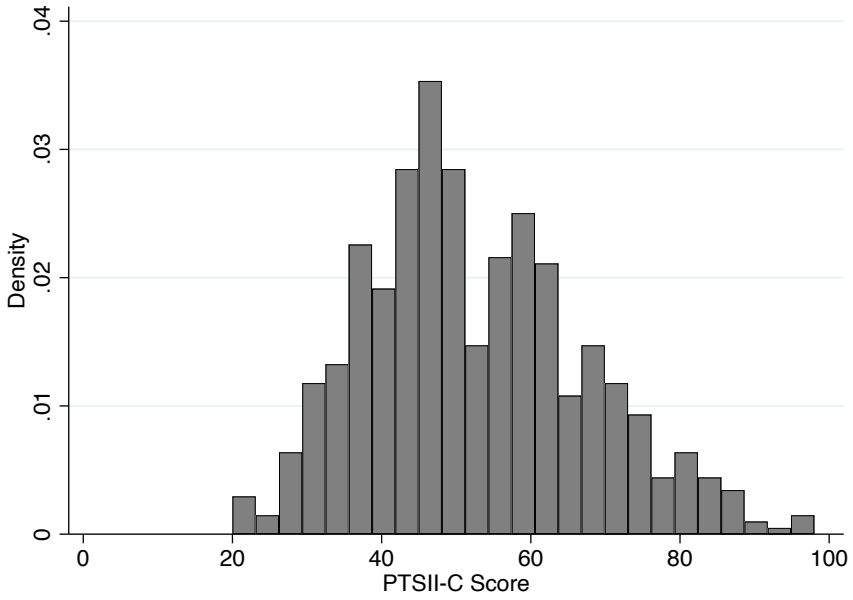
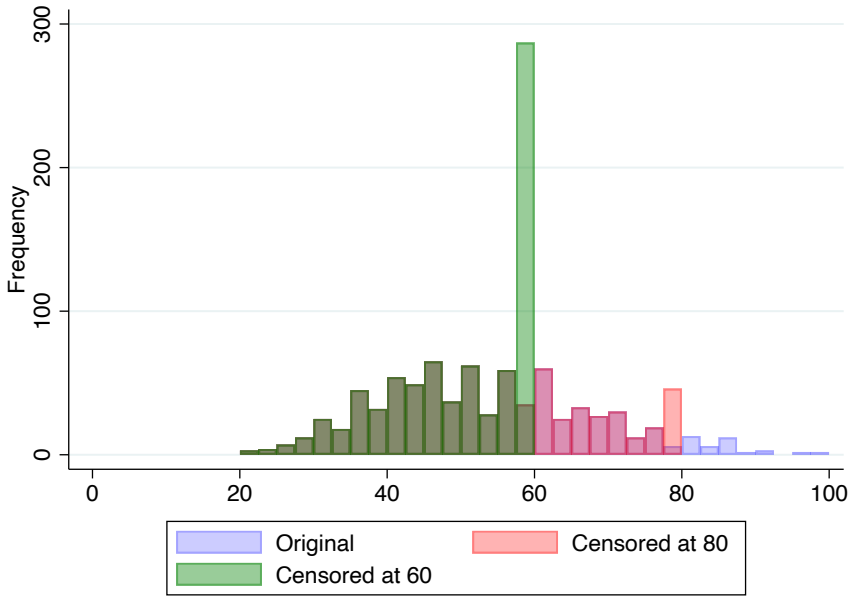
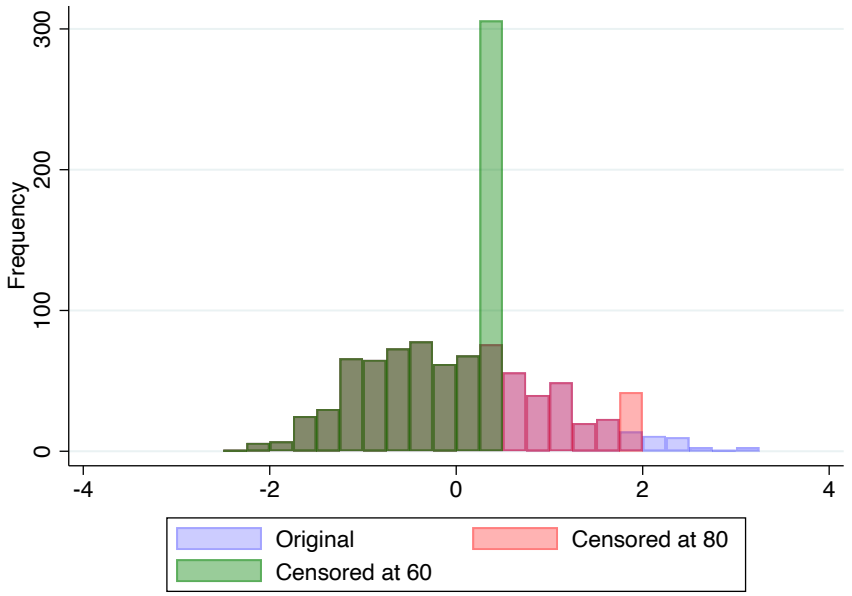


Figure 7. Distribution of PTSII-C Score



(a) Raw Score



(b) Standardized Score

Figure 8. Distributions of PTSII-C Score

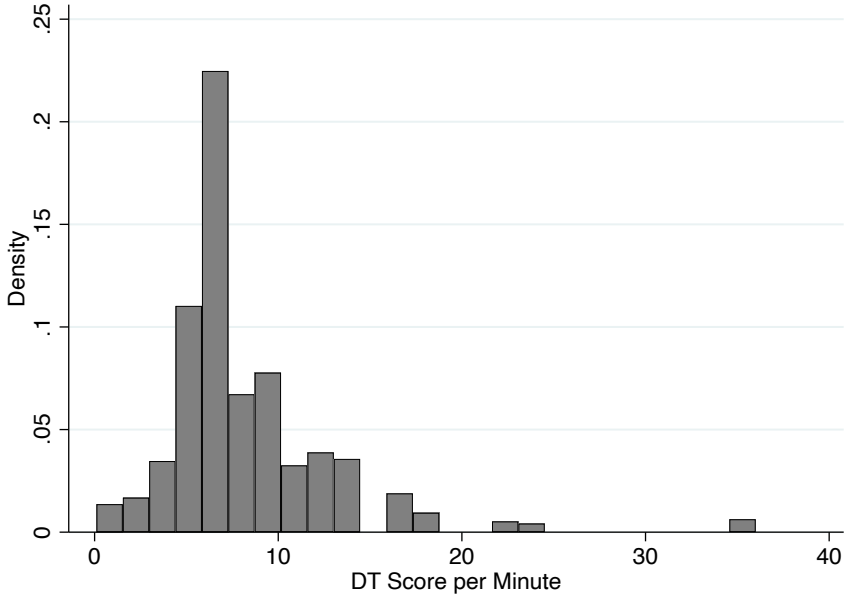


Figure 9. Distribution of DT score per minute (endline)

Table 1. Comparison between OLS and Tobit by replicating Duflo, Dupas and Kremer (2011)

	Total score								Math score				Literature score			
	(1)		(2)		(3)		(4)		(5)		(6)		(7)		(8)	
	OLS	Tobit	OLS	Tobit	OLS	Tobit	OLS	Tobit	OLS	Tobit	OLS	Tobit	OLS	Tobit	OLS	Tobit
(1) Tracking school	0.140	0.143	0.175	0.179	0.192	0.194	0.181	0.186	0.139	0.139	0.155	0.155	0.197	0.209	0.166	0.184
	(0.078)*	(0.079)*	(0.077)**	(0.078)**	(0.093)**	(0.093)**	(0.092)*	(0.093)**	(0.073)*	(0.073)*	(0.083)*	(0.082)***	(0.108)*	(0.110)*	(0.098)*	(0.102)*
(2) In bottom half of initial distribution × tracking school					-0.036	-0.034			0.039	0.039			-0.090	-0.078		
					(0.070)	(0.071)			(0.068)	(0.068)			(0.048)	(0.077)		
(3) In bottom quarter × tracking school							-0.044	-0.042			0.011	0.011			-0.082	-0.064
							(0.079)	(0.081)			(0.085)	(0.085)			(0.079)	(0.082)
(4) In second-to-bottom quarter × tracking school							-0.014	-0.018			0.025	0.025			-0.043	-0.051
							(0.068)	(0.068)			(0.076)	(0.076)			(0.068)	(0.072)
(5) In top quarter × tracking school							0.027	0.022			-0.025	-0.025			0.065	0.052
							(0.076)	(0.076)			(0.067)	(0.067)			(0.082)	(0.084)
(6) Assigned to contract teacher			0.181	0.185	0.180	0.184	0.180	0.184	0.160	0.160	0.160	0.160	0.160	0.172	0.160	0.172
			(0.038)***	(0.038)***	(0.038)***	(0.038)***	(0.038)***	(0.038)***	(0.038)***	(0.037)***	(0.038)***	(0.037)***	(0.038)***	(0.039)***	(0.038)***	(0.039)***
Individual controls	No		Yes		Yes		Yes		Yes		Yes		Yes		Yes	
Observations	5,795		5,281		5,281		5,281		5,282		5,282		5,282		5,282	

Notes: This table shows the replicated results of Table 2 (Panel A) in Duflo, Dupas and Kremer (2011).

Table 2. Comparison between OLS, Tobit, and Censored LAD by replicating Duflo, Hanna and Ryan (2012)

	Post-test								
	Math score			Language score			Total score		
	(1) OLS	(2) Tobit	(3) LAD	(4) OLS	(5) Tobit	(6) LAD	(7) OLS	(8) Tobit	(9) LAD
Treatment	0.210*	0.271*	0.215***	0.158*	0.163*	0.067*	0.173***	0.179**	0.141***
	(0.125)	(0.147)	(0.064)	(0.084)	(0.085)	(0.052)	(0.087)	(0.870)	(0.043)

Notes: This table shows the replicated results of Table 9 (Panel A) in Duflo, Hanna and Ryan (2012).

Table 3. Comparison between OLS, Tobit, and Censored LAD by replicating Duflo, Dupas and Kremer (2015)

	Total score			Math score			Literature score		
	(1) OLS	(2) Tobit	(3) LAD	(4) OLS	(5) Tobit	(6) LAD	(7) OLS	(8) Tobit	(9) LAD
Basic ETP	0.142	0.140***	0.152***	0.133	0.142***	0.219**	0.123	0.122***	0.045**
	(0.098)	(0.031)	(0.034)	(0.083)	(0.033)	(0.118)	(0.106)	(0.033)	(0.028)
ETP + SBM	0.196**	0.193***	0.239***	0.214***	0.224***	0.439***	0.141	0.148***	0.120***
	(0.098)	(0.032)	(0.041)	(0.078)	(0.034)	(0.056)	(0.103)	(0.034)	(0.029)

Notes: This table shows the replicated results of Table 2 (Panel A) in Duflo, Dupas and Kremer (2015).

Table 4. Comparison between OLS, Tobit, and Censored LAD by replicating Pradhan et al. (2014)

	Grant, G			Election, E			Linkage, L			Training, T			L+E			L+T			T+E		
	OLS	Tobit	LAD	OLS	Tobit	LAD	OLS	Tobit	LAD	OLS	Tobit	LAD	OLS	Tobit	LAD	OLS	Tobit	LAD	OLS	Tobit	LAD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)
Panel B. Language test scores																					
Average	0.129***	0.129***	0.058**	0.053**	0.050**	0.032	0.173***	0.164***	0.095***	-0.042**	-0.035	0.016	0.234***	0.223***	0.135***	0.134***	0.126***	0.097***	0.015	0.028	0.023
	(0.033)	(0.033)	(0.027)	(0.023)	(0.023)	(0.020)	(0.023)	(0.023)	(0.020)	(0.020)	(0.023)	(0.017)	(0.032)	(0.034)	(0.029)	(0.031)	(0.032)	(0.033)	(0.031)	(0.034)	(0.035)
Panel C. Math test scores																					
Average	-0.015	-0.015	-0.032	-0.008	-0.005	0.000	0.070***	0.071***	0.045*	-0.029	-0.022	0.000	0.061**	0.060**	0.070**	0.040	0.043	0.045	-0.036	-0.018	0.000
	(0.031)	(0.032)	(0.039)	(0.020)	(0.021)	(0.027)	(0.020)	(0.021)	(0.026)	(0.020)	(0.021)	(0.027)	(0.028)	(0.029)	(0.034)	(0.028)	(0.029)	(0.036)	(0.027)	(0.029)	(0.033)

Notes: This table shows the replicated results of Table 5 (Panel B,C) in Pradhan et al. (2014).

Table 5. Summary Statistics

Dependent Variable	Treatment	Control	N
PTSII-C	34.665	39.040	787
	[10.603]	[15.508]	
DT Score[a]	47.419	47.291	663
	[15.608]	[16.555]	
DT Time[a]	9.879	9.960	663
	[0.918]	[0.295]	
DT Score per min[a]	4.894	4.757	663
	[1.943]	[1.693]	
PSC GPA[b]	3.514	3.705	334
	[1.529]	[1.227]	

Notes: Standard deviations are shown in brackets. Asymptotic standard errors are shown in parentheses and are clustered at the school level. The sample size is different from Table 1 of Sawada et al. (2020), because we only focus on pupils with both baseline and endline records for each outcomes, except for PSC.
[a]: DT stands for math Diagnostic Test. DT Score per min stands for math Diagnostic Test scores per minute..
[b]: PSC GPA means the Point Average of the Primary School Certificate Grade. The letter grades from A+ to A, A-, B, C, D, and F are assigned: if the score is in the range of 80 to 100, the letter grade is an A+; if 70 to 79, it is an A; if 60 to 69, it is an A-; if 50 to 59, it is a B; if 40 to 49, it is a C; if 33 to 39, it is a D; and if below 33, it is an F. GPA is calculated as 5 if A+; 4 if A; 3.5 if A-; 3 if B; 2 if C; 1 if D; and 0 if F, following the Bangladesh government.

Table 6. Impact of Kumon on Students' Learning Outcomes

Dependent Variable	PTSII-C Score[a]	DT Score[b]	DT Time[b]	DT Score per min[b]
	(1)	(2)	(3)	(4)
Treatment	1.212***	0.501**	-2.122***	2.073***
	(0.292)	(0.226)	(0.544)	(0.570)
Constant	0.679***	0.521***	-0.881***	0.839***
	(0.212)	(0.142)	(0.227)	(0.158)
Num of Obs.	787	663	663	663
R-squared	0.193	0.048	0.182	0.168

Notes: This is from Panel C of Table E2 of Sawada et al. (2020), which focuses on the Difference-in-Differences specification. Asymptotic standard errors based on testing the hypotheses that the differences between the treatment and control is zero are shown in parentheses and are clustered at the school level. The superscripts, ***, **, *, denote the statistical significance obtained by clustered wild bootstrap-t procedures at the 1 percent, 5 percent, and 10 percent level, respectively.
[a]: PTSII-C Score stands for the math proficiency test scores.
[b]: DT stands for math Diagnostic Test. We use three outcomes of DT for measuring cognitive abilities: DT score, DT time, and DT score per minute (DT scores per min).

Table 7. Impact of Kumon on Students' Learning Outcomes

Dependent Variable	DT Score[a]	
	OLS	Tobit
	(1)	(2)
Treatment	0.490***	0.563***
	(0.137)	(0.175)
Constant	0.600***	0.616***
	(0.104)	(0.111)
N	663	663
R-squared[b]	0.095	0.045

Notes: This is from Table 2 of Sawada et al. (2020). Asymptotic standard errors based on testing the hypotheses that the differences between the treatment and control is zero are shown in parentheses and are clustered at the school level. We focus on pupils with both baseline and endline records for each outcomes. In addition, we omit the observations with wrong level DT from the analysis on DT. All variables are standardized. The superscripts, ***, **, *, denote the statistical significance obtained by (i) clustered wild bootstrap-t procedures for the OLS and (ii) (standard) clustered for the Tobit model at the 1 percent, 5 percent, and 10 percent level, respectively.
[a]: DT stands for math Diagnostic Test.
[b]: For the Tobit model, this corresponds to the pseudo-R-squared.

Table 8. Prediction of PSC Results

Independent Variable:	DT Score[a]	DT Time[a]	DT Score Per Min[a]
	(1)	(2)	(3)
Coefficient	0.152*	0.006	0.165**
	(0.080)	(0.084)	(0.082)
Constant	3.558***	3.582***	3.567***
	(0.075)	(0.075)	(0.075)
Num of Obs.	367	367	367
R-squared	0.010	0.000	0.011

Notes: Asymptotic clustered standard errors are shown in parentheses. The sample size is different from Table 1 of Sawada et al. (2020), because we only focus on pupils with both baseline and endline records for each outcomes. In addition, we omit the observations with wrong level DT from the analysis on DT. All variables are standardized. The superscripts, ***, **, *, denote the statistical significance at the 1 percent, 5 percent, and 10 percent level, respectively.
[a]: DT stands for math Diagnostic Test. DT Score per min stands for math Diagnostic Test scores per minute.

Table 9. Comparison of Estimates between Censored and Non-censored Outcomes

Dependent Variable	PTSII-C Score[a]		
	Original	Censored at 80	Censored at 60
	(1)	(2)	(3)
Panel A: OLS Estimates			
A. Treatment	0.754***	0.729***	0.514***
	(0.171)	(0.165)	(0.107)
Constant	-0.372***	-0.379***	-0.471***
	(0.086)	(0.084)	(0.060)
Num of Obs.	787	787	787
Panel B: Tobit Correction			
B. Treatment		0.750***	0.754***
		(0.172)	(0.170)
Constant		-0.370***	-0.396***
		(0.087)	(0.087)
Num of Obs.		787	787
Panel C: p-values of Statistical Tests			
I. Difference among OLS Estimates		A(1) = A(2)	A(1) = A(3)
		0.022**	0.001***
II. Difference between Specifications		A(2) = B(2)	A(3) = B(3)
		0.047**	0.001***
III. Difference between the Original Value		A(1) = B(2)	A(1) = B(3)
and the Corrected Values		0.344	0.999

Notes: Asymptotic standard errors are shown in parentheses and are clustered at the school level. The superscripts, ***, **, *, denote the statistical significance obtained by clustered wild bootstrap-t procedures for Panel A and ones by clustered for Panel B and C, at the 1 percent, 5 percent, and 10 percent level, respectively.
[a]: PTSII-C Score stands for the math proficiency test scores.