International Journal of Education and Development using Information and Communication Technology (IJEDICT), 2018, Vol. 14, Issue 3, pp. 93-112

Can Khan Academy e-learning video tutorials improve mathematics achievement in Sri Lanka?

Bilesha Weeraratne
Institute of Policy Studies, Sri Lanka

Brian Chin
Asian Development Bank, Philippines

ABSTRACT

This study evaluates the impact of Khan Academy (KA) video tutorials in a blended learning environment for ninth grade students in Sri Lanka. The 632 treatment group students followed KA during 2-3 out of 5 time slots assigned for mathematics teaching per week. Students in the control group followed the regular class room environment with a mathematics teacher during all 5 time slots. The impact of KA is evaluated based on students' mathematics achievements in raw, standardized, and Item Response Theory adjusted scaled test scores using propensity score matching. On average KA resulted in increasing students' raw and scaled test scores by 3.77 and 3.15 percentage points, respectively, and standardized test scores by 0.20 standard deviations above the mean test score. The evaluation provides initial evidence that the use of KA would help Sri Lankan students in mathematics. The findings are robust to alternative specifications, where school and teacher inputs are controlled for in addition to students' characteristics, and when estimation is limited to the subsample of students that had a common mathematics teacher across the treatment and the control groups.

Keywords*: Khan Academy, e-learning, Sri Lanka, impact evaluation*

1. INTRODUCTION

Sri Lanka, since independence in 1948, has made significant investments in education and have reached impressive levels in indicators such as adult literacy rate (93.2 % in 2016) and primary net enrollment rate (95.9 % in 2016) (Central Bank of Sri Lanka 2016). Nonetheless, there are notable issues such as mismatch between the education system and skills required in the labor market, disparities in access to quality education, and the absence of a linkage between secondary and tertiary education, and also between general and technical education and vocational training (Kim 2012). Moreover, given the high level of expenditure borne by the Government, students' low performance in mathematics at the General Certificate in Education Ordinary Level (GCE O/L) examination deserves serious attention (Department of National Planning 2013). Specifically, with the cutoff threshold for passing mathematics set at 35 marks out of 100, during 1999 to 2004 on average 60 percent of the students obtained less than 35 marks and failed mathematics (Wijewardene 2010). From 2005 to 2010 (including both years), the average mathematics failure rate was 50.31 percent (Department of Examination 2011a,b). Though lower than previous years, in 2017, 37.19 percent students have failed in mathematics in the GCE O/L examination (Hiru News 2017). Such high failure rates are amidst the notable local culture of increasing household expenditure of time and money on supplementary 'tuition' classes on top of government expenditure on free education (Kim 2012). The average monthly household expenditure on education has grown from around 3.9 percent of non-food expenditure in 1981-82 to 5.7 percent in 2016 (Department of Census and Statistics 2018). Average per person monthly expenditure on tuition was LKR 262.36 in 2016, which has experienced a 130 percent increase

over the 2009/10 figure of LKR 114.48 (Department of Census and Statistics 2011, 2018). Above statistics indicate that continuation of existing programs may not be sufficient to address issues concerning high failure rates in mathematics in Sri Lanka, and there is a growing need for "a different approach to teaching math[ematic] concepts, which is distinct from the traditional approach to learning and teaching mathematics" (Papadakis, Kalogiannakis, and Zaranis, 2016, p.250; Papadakis, Kalogiannakis, and Zaranis, 2017, p.377).

In this context, we evaluate the impact of an intervention that used Khan Academy (KA)[1] – a U.S. based not-for-profit repository of pre-recorded video tutorials, where an instructor presents mathematics concepts in 7-14 minutes on an electronic blackboard, for free of charge. KA was used in a blended learning environment in mathematics for a sample of ninth grade Sinhala[2] medium public-school students in Sri Lanka in 2014. The impact evaluation results are aimed to gain a better understanding about the impact of adopting e-learning systems on students' mathematics achievement and to gain valuable lessons and experience about implementation of same in Sri Lanka. The intervention was funded by the ADB and was implemented with the support of Ministry of Education (MoE) Sri Lanka and the National Institute of Education (NIE). The impact evaluation applies propensity score matching (PSM) on data from 562 students who were exposed to KA and 514 students who were not, spread across 20 schools in three provinces in Sri Lanka.

2. BACKGROUND

2.1 Mathematics in Sri Lanka

Among many subject students learns at school mathematics is an important one (Papadakis, Kalogiannakis, and Zaranis, 2017). This significance in mathematics is highlighted alike in education research and policy in Sri Lanka. For instance, Aturupane et al. (2011, pp.9) note that "high quality mathematics education will ensure that students develop the skills that are essential not only in science and technology, but also in everyday life and the workplace". Similarly, the education policy in Sri Lanka introduces mathematics as early as the first grade and it remains a core subject until the 11th grade. Moreover, mathematics is one of the three compulsory subjects that a student needs to pass to qualify the GCE O/L. However, in recent years the failure rates of mathematics at the GCE O/L examination in Sri Lanka has been concerning due to its implications such as dropping out of school and lower labor market outcomes later in life.

2.2 Tracking international mathematics achievement

Due to the positive correlation of education (and broader human capital development) with future labor market outcomes, many countries are interested in how students perform in mathematics at various stages of schooling. Hence, many studies track the progress of students' achievements in mathematics. Trends in International Mathematics and Science Study (TIMSS) is one such international assessment and Programme for International Student Assessment (PISA) is another. The TIMSS 2012 report identifies that schools that performed well in the assessments were likely to have better working conditions and facilities as well as more instructional materials, which included computers and technological support (IEA 2012). Similarly, in Singapore, which ranked second in PISA 2012, school principals have reported that instruction is not hindered by a shortage of resources such as computers, internet connectivity, and computer software (OECD 2013).

[1] http://www.khanacademy.org
[2] Sinhala is one of the major official and national languages of Sri Lanka.

2.3 Effectiveness of e-learning

The traditional method of teaching mathematics in Sri Lanka involves overreliance on teachers and textbooks. However, many developed and developing countries are moving towards integrating alternative teaching methods using ICT. For instance, in recent years in countries such as USA, Thailand, India, and Turkey there have been an increase in the use of computers and technology to teach mathematics (Zaranis, Kalogiannakis, and Papadakis, 2013). Nevertheless, when technology-based interventions can improve students' achievement, "it is not always clear that they are the most cost-effective means of doing so" (Dundar et al. 2014, p.256). For instance, Banerjee et al. (2007) shows that compared to a teacher based remedial program, a computer aided remedial program has the capacity to double test scores, but when scaling up the former is in the region of 5-7 times more cost-effective.

2.4 Learning from e-learning

In 2012, a pilot project on e-learning was conducted in Sri Lanka by the Asian Development Bank (ADB), called My Personal Data Analysis (MPDA).[3] The impact evaluation of the MPDA program saw that the program was effective at a statistically significant level. When standardized test scores were compared (for the twelve weeks of program implementation) the treatment groups improved by approximately 6.92 percent of the possible score range compared to the control group (Kim 2012). Kim (2012) concluded that higher effects can be expected when such a program is expanded to non-English speaking students for two reasons. First, students with English proficiency in Sri Lanka are, generally, more economically advantaged and more likely to participate in private tutoring, which would have worked as a ceiling effect in this pilot project. Second, it was unlikely to expect stable impact evaluation estimates with the limited sample size. To address these limitations, the project on KA was launched in 2014.

3. THE INTERVENTION

The intervention involved supplementing regular mathematics teaching, for a sample of ninth grade students in Sinhala medium public schools, with audio-visual tutorials in the form of YouTube videos from KA. The free audio-visual tutorials available at KA have become "one of the globe's most popular education websites" (Murphy et al. 2014, pp.3), and by 2014, KA was already used in 200 countries, while many more governments have proposed its use in their schools. A main benefit of KA is that it blended well with any teaching method and thus "…teachers did not need to change their entire teaching model to start using it" (Light and Pierson, 2014, p.117; Murphy et al. 2014; Trucano 2014). Murphy et al. (2014) report that 71 percent of KA students in their sample enjoyed using KA, and 32 percent agreed they liked mathematics more since they started using KA. In evaluating the use of KA in Chilean schools, Light and Pierson (2014) found that teachers and administrators felt that KA was effective in improving procedural skills in mathematics but not at promoting deeper mathematics learning or teaching difficult concepts.

3.1 Khan Academy in Sri Lanka

For the KA project in Sri Lanka, 115 subtopics[4] each with KA audio-video tutorials applicable to

[3] MPDA program was computer software in mathematics tailored for individual students, which provided a series of mathematics problems at various levels of content knowledge to meet each student's exact needs.
[4] Subtopics covered the content domains of Number, Algebra, Geometry, and Measurement.

the Sri Lankan ninth grade mathematics curriculum content were identified. For which, 55 existing KA videos were translated into Sinhala and another 80 new videos were developed for specific sections of the syllabus that were not available in the original KA repository. The combination of translated and newly produced material was developed by the NIE to ensure that the entire ninth grade mathematics syllabus was covered in the intervention.

3.2 Sampling

The sample of students for the intervention were selected based on a three-stage sampling procedure where initially three provinces were chosen, out of which 20 Sinhala medium public schools were selected and finally two Grade 9 classes were selected from each school for the treatment and control groups. Due to operational reasons, the entire sampling scheme was purposive (see Appendix A for details).

Students in the control group followed the ninth-grade mathematics textbook[5] and related instructions in the regular class room environment with a mathematics teacher during all 5 days of the week. The school timetable allocates at minimum 5 periods of 40 minutes each for mathematics teaching per week. Students in the treatment group followed the same course of education as the control group, but 2-3 of the 5 time slots assigned for mathematics teaching were substituted with self-study tutorials delivered through the computer where students listened and watched content.

Prior to the implementation of the intervention, mathematics teachers of treatment group classes and the ICT teachers of the 20 selected schools were familiarized about the program through a comprehensive training workshop.

3.3 Testing

In addition to blended learning, the intervention also involved testing treatment and control group students at three points in time: May 2014 (Pre-Test), October 2014 (Post Test-1), and December 2014 (Post Test-2), and concurrent surveys were conducted to gather socioeconomic and other background information of students, teachers, and schools. The intervention involved 562 ninth grade students in the treatment group and 514 ninth grade students in the control group at the Pre-Test stage, 632 and 613 in the treatment and control group students respectively in Post Test-1, and 527 and 529 treatment and control group students respectively in Post Test-2.

The three tests administered to students involved 30 test items each.[6] The Pre-Test and Post Test-2 covered all second and third academic term ninth grade mathematics syllabus content, while Post Test-1 covered only second academic term content. However, even though introduction of the use of a calculator is within the scope of the Grade 9 syllabus, students in many schools did not have calculators and thus were not familiar in using one. As such, one item from Post Test-1 and two items from Post Test-2, which involved the use of a calculator, were dropped. In Pre-Test, the average raw test scores for the full sample was 9.74 (S.D. 3.93) out of a maximum of 30, while the corresponding values for the treatment and control groups were 9.74 (S.D. 3.77) and 10.05 (S.D. 4.04), respectively. The highest average raw scores were reported in Post Test-1, where the treatment and control group averages were 12.67 (S.D. 4.95) and 12.07 (S.D. 5.28), respectively. In Post Test-2, the treatment group average raw score was 10.64 (S.D. 4.14), while the control group mean was 10.24 (S.D. 4.25).

[5] Textbooks are provided free of charge by the government.
[6] Since testing for KA intervention was conducted in addition to regular term examinations, the possibility of 'teaching for the test' is eliminated.

4. METHODOLOGY

4.1 Scores

One of the limitations of raw scores is that it does not facilitate meaningful comparison of test scores across tests. For in depth comparison of test scores, it is important to know more information, such as the mean and the number of standard deviations one's score is above or below the mean. Therefore, as an alternative to raw scores, standardized test scores are considered. Standardized test score *(Z_i)* is calculated as shown in Equation (1), where X_i is student *i*'s test score in a given test. $\bar{X}$ is the mean test score of the sample of students, and *SD*(*X*) is the standard deviation of test scores for the entire sample. Standardized test scores facilitate comparison of raw scores that come from very different sources such as Pre-Test and Post Tests and facilitate to analyze students' performance relative to the sample of students who took the test.

$$Z_i = (X_i - \bar{X})/(SD(X)) \quad (1)$$

Even though standardized scores are a better comparison across tests than raw scores, they both fail to "account for differences between the individual questions on the test" (Naseer et al. 2010, pp. 675). For instance, consider two students who obtained equal (*n)* marks on Pre-Test "and correctly answered the same set of *n − 1* questions. The only difference was that student A correctly answered a one-digit addition whereas student B got that question wrong but correctly answered a more complicated two-digit addition with carry" (Naseer et al. 2010, pp. 688). These subtle details of testing and scoring are not depicted in raw and standardized scores. Such differences in test items are important in the context of testing done for KA intervention. Specifically, as mentioned before, the Pre-Test questions were based on second and third term syllabus content and the test was administered at the beginning of second term, before most of the content was taught to students. On the contrary, Post Test-1 covered only the content of the second academic term and was administered well after the second term subject matter was taught in schools, whereas Post Test-2 was like the Pre-Test in having included subject matter relevant to both second and third term syllabus but was administered right after learning. As such, the varying levels of difficulty of test items relative to the time of testing calls for comparison of a more sophisticated score such as scaled scores.

Scaled scores used in this evaluation are Item Response Theory (IRT) adjusted aggregate scores. "IRT models student ability using question level performance instead of aggregate test level performance. Instead of assuming all questions contribute equally to our understanding of a student's abilities, IRT provides a more nuanced view" about a student depending on the information provided by each question (Companioni 2012). To account for the difficulty of each question and the ability of each question to discriminate between students with high- and low-ability, a Two Parameter Logistic (2PL) IRT model is adopted, as shown in Equation (2).

$$P(X_{ij} = 1 | \theta_j, \alpha_i, \beta_i)) = exp[\alpha_i(\theta_j - \beta_i)]/(1 + exp[\alpha_i(\theta_j - \beta_i)]) \quad (2)$$

In Equation 2, θ_j refers to the ability of student *j*, α_i is the discrimination of item *i*, and β_i is the difficulty of item *i.* Based on the above 2PL model the probability of a correct response by an individual to each test item is estimated and the scaled score for student *j* is arrived by summing the probability of an item to be correct $[P(X_{ij} = 1)]$ over all 30 items as shown in Equation 3.

$$\text{Scaled Score of student } j = \sum^{30}[P(X_{ij} = 1)] \quad (3)$$

4.2 Propensity Score Matching

The impact evaluation adopts a PSM methodology developed by Rosenbaum and Rubin (1983). PSM controls for possible selection bias and makes the treatment and control groups more comparable based on observable characteristics. PSM estimates the potential outcome for each student - the outcome if the student was assigned to the other group, by using an average of the outcomes of similar students that were in alternative group. Similarity between subjects is based on estimated treatment probabilities, which are propensity scores.

PSM is successful in treatment evaluation when those in treatment and control groups can be matched on observable characteristics, and the key identifying assumption is outcomes are independent of treatment assignment, given observable characteristics. This unconfoundedness assumption, which means that once the differences in observed pretreatment variables are controlled for, the biases between the treatment and control groups are removed, is expressed as $Y_0 \perp D \,|P(X)$. Here Y_0 is the outcome of interest, in the untreated states, *D* is the treatment indicator, *X* is a vector of observable characteristics, while *P(X)* is the propensity score. Successful PSM also requires to meet the common support or overlap condition, which rules out the possibility of perfect predictability of *D* given *X* expressed as $0 < P(D = 1|X) < 1$. The overlap condition ensures that student "with the same *X* values have a positive probability of being both participants and non-participants" (Heckman, LaLonde, and Smith, 1999, pp. 1920).

5. DATA AND ESTIMATION

Data for the analysis is constructed by combining test scores with survey data collected for the intervention. The outcome of analysis in PSM exercise is the difference in score between Pre-Test and Post Test-1 as well as Pre-Test and Post Test-2.[7] Table 1 depicts summary statistics for the various outcome variables used by unmatched and matched samples. The outcome variables reported are of differences in raw, standard, and scaled test scores across Pre-Test and Post Test-1, and Pre-Test and Post Test-2. As seen, in the unmatched sample, mean test scores of the treatment group students are significantly higher than that of control group students.

In PSM, high quality matching is critical, as students across the two groups are not matched on all covariates, but on propensity scores. The choice of independent variables used in the matching equation are guided by economic theory and evidence, where characteristics of student, their family background, schools, and teachers are incorporated (Fryer 2017). The matching equation uses students' mathematic ability indicated by the Pre-Test scaled scores and confidence in mathematics, age indicated by birth year, self-reported expectation about future education, computer savviness indicated by students' attitude towards using computers, and number of days attending tuition classes. The family socio-economic background is indicated by variables such as receipt of *Samurdhi* benefits[8], usage of English language and computers, and parents' level of education. The variables included influence simultaneously the participation decision and the outcome variable, while the choice of variables ensure some randomness that persons with identical characteristics can be observed in both states. Matching adopted a nearest neighbor matching algorithm where each student from the treatment group is matched with a student from the control group that has the nearest propensity score, on "the region of common support defined as the maximum of min[imum]s, and the minimum of max[imum]s" (Himaz 2008, pp.1848).

[7] For the PSM analysis, test scores are converted to percentages.

[8] *Samurdhi* is a cash transfer program to low income households.

Matching is considered high quality if the distribution of covariates of the relevant variables is balanced across the treatment and control groups and relevant assumptions are validated. As detailed in Appendix B1, covariates are balanced across the two groups after matching. Moreover, the assumptions on unconfoundedness and overlap condition are validated in Appendix B2 and B3, respectively.

Table 1: Summary Statistics for Outcome Variables by Unmatched and Matched Samples

	Unmatched Sample				Matched Sample			
	Treatment (1)	Control (2)	t-stat (p-value) between (1) and (2) (3)		Treatment (4)	Control (5)	t-stat (p-value) between (4) and (5) (6)	
Raw score diff Pre to Post1	11.98	9.06	-3.41	***	12.70	9.97	-2.50	*
	(0.62)	(0.59)	(0.00)		(0.00)	(0.78)	(0.01)	
Raw score diff Pre to Post2	5.33	3.65	-2.01	**	4.37	3.36	-1.03	
	(0.60)	(0.59)	(0.04)		(0.67)	(0.71)	(0.30)	
Std score diff Pre to Post1	0.10	-0.09	-3.48	***	0.12	-0.05	-2.35	*
	(0.04)	(0.04)	(0.00)		(0.05)	(0.05)	(0.02)	
Std score diff Pre to Post2	0.07	-0.03	-1.86	*	0.00	-0.06	-1.00	
	(0.04)	(0.04)	(0.06)		(0.05)	(0.05)	(0.32)	
Scaled score diff Pre to Post1	11.13	8.90	-3.34	***	12.04	9.60	-2.83	***
	(0.46)	(0.48)	(0.00)		(0.61)	(0.61)	(0.00)	
Scaled score diff Pre to Post2	5.12	3.52	-2.59	***	4.38	3.34	-1.47	
	(0.44)	(0.44)	(0.01)		(0.48)	(0.52)	(0.14)	

Notes: For columns (1), (2), (4) and (5), standard errors in parentheses. For columns (3) and (6), p-values in parentheses. * p<0.10, ** p<0.05, *** p<0.01.
Source: Author's calculations based on field data.

6. RESULTS AND DISCUSSION

6.1 Main models – matching on students' characteristics.

With validation of the appropriateness of PSM methodology, the analysis focuses on two treatment effects: Average Treatment Effect (ATE) and Average Treatment on the Treated (ATT). ATE considers the average increase in test scores if the all students in the control group were given the opportunity to use KA. On the contrary, ATT focuses on the average effect of treatment on those subjects who received the treatment. Impact evaluation estimates based on PSM are reported in Table 2. The left panel of Table 2 corresponds to impact evaluation at the end of Post Test-1 (5 months after inception) and right-side panel corresponds to that of Post Test-2 (7 months after inception). The three columns in each panel correspond to the impact estimates based on raw, standardized, and scaled scores.

As seen in column (1), when raw scores are considered as the outcome, the ATE of KA is 2.54. In terms of standardized scores at the end of Post Test-1 KA program has resulted in increasing test scores of a randomly selected student by 0.15 standard deviations above the average. The ATE estimates for scaled scores and estimates for Post Test-2 are not statistically significant. The absence of statistically significant estimate indicates that the impact after 7 months is not different from zero or no impact. When ATT is considered, none of the models produced statistically significant impact estimates of the KA program although the magnitudes and directions are similar to ATE.

Table 2: Treatment Effect Estimates: Main Models

	Post Test-1 (after 5 months)			Post Test-2 (after 7 months)		
Period	Raw	Std	Scaled	Raw	Std	Scaled
Outcome	(1)	(2)	(3)	(4)	(5)	(6)
ATE	2.54 **	0.15 **	1.45	-0.01	-0.01	0.73
	(1.16)	(0.07)	(0.93)	(1.17)	(0.08)	(0.76)
ATT	1.61	0.09	1.21	-0.58	-0.05	0.38
	(1.35)	(0.08)	(1.07)	(1.44)	(0.10)	(0.90)
Obs	788	788	788	773	773	773

Notes: Standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01.
Source: Author's estimates.

6.2 Robustness Check-1 – adding teacher and school level characteristics

The above estimated models are based on matching on students' characteristics. However, as widely noted in literature, in addition to student characteristics, the education production function also includes teacher and school level inputs (Ronfeldt et al. 2015). Hence, to test if above findings are robust when school and teacher level characteristics are considered in the matching equations, additional controls: school size depicted by number of students in school and an indicator if teacher has participated in any computer training programs, are introduced in Robustness Check-1. The top half of Table 3 depicts impact estimates under this robustness check.

When the matching equation controls for these additional teacher and school level information, the ATE estimates are statistically significant for all three types of scores for Post Test-1. At the end of Post Test-1, KA has resulted in increasing raw test scores by 2.72 percentage points. This finding shows that the previous finding in terms of ATE of main model is robust to alternative

specification where school and teacher level inputs are also considered. Compared to the impact of 2.54 percentage points estimated previously, with additional controls the magnitude of the impact has increased by 0.18 percentage points.

Noticeably, when controlled for teacher and school characteristics, previously significant standardized test scores of Post Test-1 has remained almost the same and statistically significant. Moreover, scaled scores at the end of Post Test-1 has now produced statistically significant impact estimates. Specifically, KA has resulted in increasing scaled scores for a randomly selected student in this sample by 2.34 percentage points. As in the main models, none of the ATT estimates for Robustness Check-1 are statistically significant.

Table 3: Robustness Checks

	Post Test-1			Post Test-2		
Period	Raw	Std	Scaled	Raw	Std	Scaled
Outcome	(1)	(2)	(3)	(4)	(5)	(6)
Robustness Check-1[a]						
ATE	2.72 **	0.14 **	2.34 ***	0.43	0.03	0.49
	(1.15)	(0.07)	(0.87)	(1.13)	(0.08)	(0.75)
ATT	1.72	0. 06	1.29	0.09	0.01	0.08
	(1.36)	(0.08)	(0.99)	(1.37)	(0.09)	(0.88)
Obs	744	744	744	719	719	719
Robustness Check-2[b]						
ATE	5.01 **	0.27 *	3.96 ***	1.97	0.12	2.24 *
	(2.43)	(0.15)	(1.49)	(1.96)	(0.14)	(1.31)
ATT	2.47	0.10	2.20	4.11 **	0.26 *	3.85 ***
	(2.70)	(0.16)	(1.66)	(2.03)	(0.14)	(1.38)
Obs	164	164	164	167	167	167

Notes: Standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01.
[a] Matching based on students', teachers' and school characteristics.
[b] Matching based on students', teachers' and school characteristics on the subsample with a common mathematics teacher across treatment and control groups.

6.3 Robustness Check-2 – restricting to a common mathematics teacher

In addition to school level inputs and teacher's characteristics, teaching style also may contribute to students' educational outcomes (Catalán et al. 2018). As the intervention involved training teachers on KA, one may argue that teachers in the treatment group are better trained at teaching. Moreover, as explained in Appendix A, mathematics teachers self-selected into treatment group. Hence, the finding discussed so far can be criticized for being driven by better teaching in the treatment groups. To counter this argument, a second robustness test limits estimation to five schools where a common teacher taught mathematics to both groups. As in the previous robustness test, here also all student, teacher and school level information are used in the matching equation. As evident in the bottom panel of Table 3, all ATE estimates for Post Test-1 are statistically significant. The ATE on Post Test-1 raw test scores for the sample students with a common mathematics teacher is a 5.01 percentage point increase. This is the largest impact estimate of KA program and approximately double the hitherto estimates for Post

Test-1 raw test scores. Similarly, in terms of standardized test scores and scaled scores also the estimates under Robustness Check-2 are the largest. Specifically, at Post Test-1 the ATE of KA is increase in standardized test score by 0.27 standard deviations above the average and a 3.96 percentage point increase in scaled test scores. Additionally, the impact estimates for scaled scores at Post Test-2 are also significant, which shows that after 7 months of implementation the KA program has a causal effect of increasing scaled score by 2.24 percentage points. In terms of ATT, all estimates of Post Test-2 are statistically significant at conventional levels. Hence, at the end of 7 months on average a randomly selected student from the treatment group with common teacher would experience 4.11 percentage points increase in raw test scores, 0.26 standard deviations above average in standard scores, and 3.85 percentage points increase in scaled scores. However, given the timing of tests, findings based on Post Test-1 may reflect achievement on material just taught, while statistically significant results based on Post Test-2 are the more appropriate findings for this impact analysis.

The above quantitative findings of the study are also supported by the qualitative responses provided by students and teachers. For instance, at the end of Post Test-1 over 72 percent of the students in the Treatment group agreed at varied levels that the program should be continued, while 8 percent were indifferent. Only 7 percent of the students that were exposed to KA program thought the program should not continue. Likewise, at the end of Post Test-2 also most students thought that the program should be continued, while only 6 percent disagreed. Similarly, at the end of the program (Post Test-2) nearly 72 percent of the treatment group students agreed that this program has improved their mathematics, while 12 percent were indifferent and only around 4 percent disagreed. Additionally, at the end of the program over 74 percent agreed that KA would further improve their mathematics, while 11 percent were indifferent. Only 4 percent disagreed that the continuation of the program would further improve their mathematics. Echoing students' sentiments, teachers also felt that KA has contributed to improve students' interest in mathematics and that KA was useful to students.

7. CONCLUSION

The impact evaluation shows a positive impact of KA on students' achievement in mathematics. The impact of KA on raw test scores ranged between an increase of 2.54-5.01 percentage points and an average of 3.77 percentage points, while the impact on standardized scores are between 0.14-0.27 standard deviations above the mean. The average increase in standardized scores is 0.20 standard deviations above the mean. In terms of scaled scores – the more superior measure, the impact of KA program ranges between 2.34- 3.96 percentage points increase in students' scaled test scores and an average increase of 3.15 percentage points. Two findings are especially important to highlight where the teacher has a large influence on students' achievement in using e-learning systems. First, the ATT achievement is higher, i.e., the impact of the KA program is greater for a randomly picked student in the treatment group students than a randomly picked student in the entire sample. As such, in scaling up KA, ATE would provide a more realistic indication of possible improvements in test scores. Second, the achievement is largest where there is a common teacher who teaches both treatment and control students. Both findings show that the willingness of teachers to participate and adopt blended learning cannot be understated, while there is a quality in these teachers that was common to higher student achievement.

Like other KA interventions and evaluations in other settings (Murphy et al. 2014, Light and Pierson 2014, Libraries without Borders 2014, Funsepa 2016), the KA model implemented in a sample of 20 public schools in Sri Lanka was not designed to replace the teacher. Rather, KA was used as a teaching tool supplementary to traditional classroom teaching (2-3 days KA and the rest 2-3 days traditional teaching). The teacher still leads all classes and students can pause,

watch and listen to the KA tutorials and interact with each other and the teacher, which improves learning outcomes, and possibly teaching better. Although teachers did not need to radically change their entire teaching model and curriculum to start using KA, teachers could engage more easily in differentiated instruction and took it as an opportunity to refresh pedagogical skills and lesson planning. KA supports blended and personalized learning where "students engage with and are engaged by the mathematics content; it also changes the way teachers and students interact with each other" (Light and Pierson 2014, pp. 114).

In conclusion, this evaluation provides initial evidence that using KA would help Sri Lankan students to overcome some of the challenges related to the high failure rate in mathematics. Nonetheless, before scaling up the introduction of e-learning video tutorials for mathematics education in Sri Lanka, more analysis is required. Future work in this area should focus on the ideal duration of such interventions and the long-term gains to students' achievement, by adopting an intervention that follows students from ninth grade until they receive their GCE O/L examination results.

REFERENCES

Aturupane, H., Dissanayake, V., Jayewardene, R., Shojo, M., and Sonnadara, U. 2011. Strengthening Mathematics Education in Sri Lanka. *World Bank, South Asia Human Development Sector, Discussion Paper Series* - Report No. 43. World Bank, Washington, D.C.

Banerjee, A.V., Cole, S., Duflo, E., and Linden., L. 2007. Remedying Education: Evidence from Two Randomized Experiments in India. *The Quarterly Journal of Economics*, Vol. 122, No. 3, pp. 1235-1264.

Catalán, Á.A., Serrano, J.S., Lucas, J.M.A., Clemente, J.A.J., and García-González, L. 2018. An Integrative Framework to Validate the Need-Supportive Teaching Style Scale (NSTSS) in Secondary Teachers through Exploratory Structural Equation Modeling. *Contemporary Educational Psychology*. Vol. 52, pp. 48-60.

Central Bank of Sri Lanka. 2016. *Annual Report 2016*. Central Bank of Sri Lanka, Colombo.

Companioni, A. R. 2012. *The Mismeasure of Students: Using Item Response Theory Instead of Traditional Grading to Assess Student Proficiency.* Available at https://medium.com/knerd/the-mismeasure-of-students-using-item-response-theory-instead-of-traditional-grading-to-assess-b55188707ee5 [Accessed June 5, 2018].

Department of Census and Statistics. 2011. *Household Income and Expenditure Survey 2009/10, Final Report*. Ministry of Finance and Planning, Sri Lanka.

Department of Census and Statistics. 2018. *Household Income and Expenditure Survey 2016, Final Report*. Ministry of Finance and Planning, Sri Lanka.

Department of Examination. 2011a. *Statistical Handbook 2005-2007*. Department of Examinations, Sri Lanka.

Department of Examination. 2011b. *Statistical Handbook 2008-2010*. Department of Examinations, Sri Lanka.

Department of National Planning. 2013. *Mahinda Chintana Vision for the Future; Public*

Investment Strategy 2014-2016. Ministry of Finance and Planning, Sri Lanka.

Dundar, H., Bteille, T., Riboud, M., and Deolalikar, A. 2014. *Student Learning in South Asia Challenges, Opportunities, and Policy Priorities*. Directions in Development: Human Development. Washington, DC: World Bank.

Fryer Jr, R.G. 2017. The Production of Human Capital in Developed Countries: Evidence from 196 Randomized Field Experiments. In *Handbook of Economic Field Experiments,* Vol. 2, pp. 95-322. North-Holland.

Funsepa. 2016. Assessing the use of Technology and Khan Academy to Improve Educational Outcomes in Sacatepéquez, Guatemala. Available at https://learningequality.org/media/FUNSEPA_Final_Evaluation_Report_27May2016.pdf [Accessed June 5, 2018].

Heckman, J., LaLonde, R. and Smith J. 1999. The Economics and Econometrics of Active Labor Market Programs. In O. Ashenfelter, and D. Card, eds., *Handbook of Labor Economics Vol.III*, pp. 1865-2097. Elsevier, Amsterdam.

Himaz, R. 2008. Welfare Grants and Their Impact on Child Health: The Case of Sri Lanka. *World Development*, Vol. 36, No. 10, pp. 1843-1857.

Hiru News. 2017. GCE O/L Pass Rate Increase by 7 %. Hiru News March 29, 2017. Available at http://www.hirunews.lk/158005/gce-o-l-mathematics-pass-rate-increases-by-7 [Accessed June 5, 2018].

IEA. 2012. *TIMSS 2011 Assessment*. TIMSS & PIRLS International Study Center, and International Association for the Evaluation of Educational Achievement (IEA), IEA Secretariat, Amsterdam, the Netherlands.

Kim, J. 2012. Results from Item Response Theory (IRT) Analysis of The MPDA Mathematics Program in Sri Lanka (First Term 2012). Asian Development Bank, Manila, Philippines.

Libraries without Borders. 2014. *Transforming Education: Preliminary Results from the Study Carried Out in Cameroon on the Impact of Khan Academy on Children's Academic Achievement and Cognitive Abilities.* Available at https://learningequality.org/media/Rapport-Etude-Cameroun_KL_ENG.pdf [Accessed June 5, 2018].

Light, D. and Pierson, E. 2014. Increasing Student Engagement in Math: The Use of Khan Academy in Chilean Classrooms. *International Journal of Education and Development using Information and Communication Technology*, Vol. 10, No. 2, pp. 103-119.

Murphy, R., Gallagher, L., Krumm, A., Mislevy, J., and Hafter, A. 2014. *Research on the Use of Khan Academy in Schools*. Research Brief, SRI International, Menlo Park, CA.

Naseer, M. F., Patnamb, M., and Razac, R. R. 2010. Transforming Public Schools: Impact of the CRI Program on Child Learning in Pakistan. *Economics of Education Review*, Vol. 29, No. 4, pp. 669-683.

OECD. 2013. *PISA 2012 Results: What Makes Schools Successful? Resources, Policies and Practices* (Volume IV), PISA, OECD Publishing.

Papadakis, S., Kalogiannakis, M., and Zaranis, N. 2017. Improving Mathematics Teaching in Kindergarten with Realistic Mathematical Education. *Early Childhood Education Journal*, Vol. 45, No. 3, pp. 369-378.

Papadakis, S., Kalogiannakis, M., and Zaranis, N. 2016. Comparing Tablets and PCs in Teaching Mathematics: An Attempt to Improve Mathematics Competence in Early Childhood Education. *Preschool and Primary Education*, Vol. 4, No. 2, pp. 241-253.

Ronfeldt, M., Farmer, S.O., McQueen, K., and Grissom, J.A. 2015. Teacher Collaboration in Instructional Teams and Student Achievement. *American Educational Research Journal*, Vol. 52, No. 3, pp. 475-514.

Rosenbaum, P.R. and Rubin, D.B. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, Vol. 70, No. 1, pp. 41-55.

Trucano, M. 2014. *Evaluating the Khan Academy.* EduTech. A World Bank Blog on ITC use in Education. Available at http://blogs.worldbank.org/edutech/evaluating-khan-academy [Accessed June 5, 2018].

Wijewardene, W.A. 2010. *Sri Lanka's Myth of Free Education and its Quality*. Watch Tower. Lanka Business Online. Available at http://archive.lankabusinessonline.com/news/sri-lankas-myth-of-free-education-and-its-quality/523309525 [Accessed June 5, 2018].

Zaranis, N., Kalogiannakis, M., and Papadakis, S. 2013. Using Mobile Devices for Teaching Realistic Mathematics in Kindergarten Education. *Creative Education*, Vol. 4, pp. 1-10.

Original article at: http://ijedict.dec.uwi.edu/viewarticle.php?id=2488

APPENDIX A: SAMPLING SCHEME

At the stage of sampling provinces, the availability of ICT infrastructure and spread of Sinhala medium students challenged random sampling. Hence, three provinces – Western, Southern, and Central, were purposively selected due to their relatively comparable ICT infrastructure and spread of Sinhala medium students. Within these provinces, the requirement of 20 computers for simultaneous students use necessitated the 20 schools to be purposively selected. In terms of sampling, two classes within the selected 20 schools, human resource issues necessitated purposive sampling. As such, within each school the treatment group classes were often selected based on availability of a teacher who was willing/available to attend the training workshops related to the project and was willing to undertake the extra burden of incorporating e-learning to the established teaching methodologies.

APPENDIX B: VALIDATING PSM METHODOLOGY

B1. Covariate balancing

A common approach to establish balance is the use of two sample t-test to show the absence of any significant the difference in means of covariates across the two groups after matching.

Table B1 exhibit the summary statistics for relevant variables before and after matching. Columns 1,2,4 and 5 present mean values and their standard errors in parenthesis. T-test statistics reported in columns 3 and 6 are of testing for equality of means across each group in respective samples. As seen in Column 3 of Table B1, in the unmatched sample the mean values of 11 variables are statistically significantly difference across the treatment and the control groups. In the matched sample (column 6), as expected majority of these variables have been balanced to eliminate statistically significant differences across the two means.

Table B1: Summary Statistics for conditioning variables by unmatched and matched samples

		Unmatched Sampled			Matched Sample		
		Treatment (1)	Control (2)	t-stat between (1) and (2) (p-value) (3)	Treatment (4)	Control (5)	t-stat between (4) and (5) (p-value) (6)
Pre-Test scaled score		10.14	10.38	1.51	10.44	10.54	0.55
		(0.10)	(0.11)	(0.13)	(0.14)	(0.15)	(0.58)
Post Test-1 scaled score		13.12	12.79	-1.27	13.62	13.11	-1.53
		(0.17)	(0.19)	(0.20)	(0.21)	(0.24)	(0.13)
Post Test-2 scaled score		11.06	10.62	-2.02 **	10.97	10.78	-0.75
		(0.15)	(0.16)	(0.04)	(0.17)	(0.18)	(0.45)
Birth year		2000.07	2000.07	0.0364	2000.06	2000.09	1.71
		(0.013)	(0.013)	(0.97)	(0.014)	(0.015)	(0.087)
No. days Tuition		3.57	3.55	-0.33	3.71	3.61	-0.80
		(0.06)	(0.06)	(0.74)	(0.08)	(0.08)	(0.42)
Confidence in Math		66.22	63.32	-2.43 **	67.30	65.14	-1.42
		(0.83)	(0.86)	(0.02)	(1.06)	(1.08)	(0.16)
Edu. expected	- Pass GCE	0.13	0.13	0.03	0.12	0.12	0.03
		(0.01)	(0.01)	(0.97)	(0.01)	(0.01)	(0.97)
	- Pass GCE A/L	0.10	0.06	-2.41 **	0.10	0.07	-1.59
		(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.11)
	- Vocational/ technical	0.04	0.02	-2.14 **	0.04	0.02	-2.00 **
		(0.01)	(0.01)	(0.03)	(0.01)	(0.01)	(0.05)
	- Degree	0.47	0.42	-1.67 *	0.49	0.43	-1.80 *
		(0.02)	(0.02)	(0.09)	(0.03)	(0.03)	(0.07)
	- Post-graduate	0.10	0.14	1.85 *	0.10	0.15	2.11 **
		(0.01)	(0.01)	(0.07)	(0.02)	(0.02)	(0.03)
	- I don't know	0.16	0.23	3.10 ***	0.15	0.22	2.52 **
		(0.01)	(0.02)	(0.00)	(0.02)	(0.02)	(0.01)

		Unmatched Sampled			Matched Sample		
		Treatment (1)	**Control (2)**	**t-stat between (1) and (2) (p-value) (3)**	**Treatment (4)**	**Control (5)**	**t-stat between (4) and (5) (p-value) (6)**
English use socially	- Always	0.01	0.02	1.08	0.01	0.02	0.89
		(0.00)	(0.00)	(0.28)	(0.01)	(0.01)	(0.37)
	- Sometimes	0.81	0.82	0.53	0.84	0.85	0.32
		(0.02)	(0.02)	(0.59)	(0.02)	(0.02)	(0.75)
	- Never	0.18	0.16	-0.86	0.15	0.13	-0.67
		(0.01)	(0.01)	(0.39)	(0.02)	(0.02)	(0.50)
Mother's Edu	–no schooling	0.18	0.20	1.27	0.15	0.20	1.99
		(0.02)	(0.02)	(0.20)	(0.02)	(0.02)	(0.05) **
	- GCE O/L or up to GCE A/L	0.13	0.13	0.26	0.14	0.12	-0.91
		(0.01)	(0.01)	(0.8)	(0.02)	(0.02)	(0.36)
	-GCE A/L	0.33	0.27	-2.15	0.36	0.27	-2.84
		(0.02)	(0.02)	(0.03) **	(0.02)	(0.02)	(0.01) **
Vocational/ Technical		0.24	0.24	0.05	0.22	0.25	0.81
		(-0.02)	(-0.02)	(-0.96)	(-0.02)	(-0.02)	(-0.41) ***
Province	- Central	0.3	0.32	0.46	0.29	0.33	1.28
		(-0.02)	(-0.02)	(-0.65)	(-0.02)	(-0.02)	(-0.2)
	- Southern	0.28	0.31	1.54	0.31	0.29	-0.76
		(-0.02)	(-0.02)	(-0.12)	(-0.02)	(-0.02)	(-0.45)
	- Western	0.42	0.37	-1.87 *	0.39	0.38	-0.49
		(-0.02)	(-0.02)	(-0.06)	(-0.02)	(-0.02)	(-0.62)
Samurdhi recipient	-Yes	0.19	0.21	0.68	0.22	0.18	-1.41
		(-0.01)	(-0.01)	(-0.49)	(-0.02)	(-0.02)	(-0.16)
Like computer	-Like	0.86	0.86	0.53	0.86	0.86	0.39
		(-0.01)	(-0.01)	(-0.59)	(-0.02)	(-0.02)	(-0.7)

	Unmatched Sampled			Matched Sample		
	Treatment (1)	**Control (2)**	**t-stat between (1) and (2) (p-value) (3)**	**Treatment (4)**	**Control (5)**	**t-stat between (4) and (5) (p-value) (6)**
- Indifferent	0.09	0.06	-1.68 *	0.09	0.06	-1.39
	(-0.01)	(-0.01)	(-0.09)	(-0.01)	(-0.01)	(-0.16)
-Unlike	0.04	0.06	1.24	0.03	0.05	1.19
	(-0.01)	(-0.01)	(-0.21)	-0.01	-0.01	(-0.232)
Like to learn math on computer						
-Agree a lot	0.49	0.49	0.32	0.52	0.51	-0.36
	(-0.02)	(-0.02)	(-0.74)	(-0.03)	(-0.03)	(-0.72)
- Agree a little	0.37	0.37	0.037	0.37	0.36	0.65
	(-0.02)	(-0.02)	(-0.97)	(-0.03)	(-0.03)	(-0.51)
- Neither agree nor disagree	0.09	0.07	-1.68 *	0.09	0.06	-1.4
	(-0.01)	(-0.01)	(-0.09)	(-0.01)	(-0.01)	(-0.16)
- Disagree a little	0.03	0.03	0.21	0.02	0.03	1.16
	(-0.01)	(-0.01)	(-0.83)	(-0.01)	(-0.01)	(-0.25)
- Disagree a lot	0.02	0.03	1.62	0.02	0.02	0.48
	(-0.01)	(-0.01)	(-0.1)	(-0.01)	(-0.01)	(-0.63)
Computer at home - Yes	0.49	0.46	-1.07	0.55	0.54	-0.36
	(-0.02)	(-0.02)	(-0.28)	(-0.03)	(-0.03)	(-0.72)
School size	1651.6	1639.8	-0.23	1649.5	1678.8	0.67
	(-36.4)	(-37.53)	(-0.82)	(-46.04)	(-50.3)	(-0.67)
Teacher computer trained - Yes	0.45	0.46	0.36	0.51	0.54	1.07
	(-0.25)	(-0.03)	(-0.71)	(-0.02)	(-0.02)	(-0.28)

Note: In columns (1), (2), (4) and (5): for continuous variables, the values are means and standard errors, in parentheses: and for binary variables, the values are of proportions and standard errors in parentheses. Matched sample means reported are for the matched observations used for estimation of models. *p<10, **p<0.05, ***p<0.01

B2. Unconfoundedness assumption

Propensity score matching technique relies heavily on the unconfoundedness assumption. Despite the inability to directly test unconfoundedness, available data from the intervention facilitates indirectly testing this assumption via a pseudo outcome approach. The pseudo outcomes considered here are pretreatment test scores of both groups. The Pre-Test scores are closely related to the outcome of interest (changes in test scores), while it is known as a priori to not be affected by the treatment. Given that scaled version of Pre-Test scores are included in the matching variables as a proxy for ability; we rely on standardized test scores at the Pre-Test as the pseudo outcome. In addition to ability, all other independent variables in the matching equation are the same as used in the above analyses.

Table B2: Unconfoundedness Assessment

Outcome	ATE	S.E.
Pseudo outcome (Std. Score at Pre-Test)	-0.00[a]	0.05
Std. Score at Post Test-1	0.19***	0.07
Std. Score at Post Test-2	0.14**	0.07

[a] value >0.00
*p<0.10 ** p<0.05 *** p<0.01
Source: Author's calculations based on field data.

As seen in the first row of Table B2, the ATE of the pseudo outcome is not statistically significant. This indicates that we fail to reject the null hypothesis that the pseudo causal effect is equal to zero, and that the unconfoundedness assumption is plausible. This confirms that when controlled for observed pretreatment variables, the bias between the treatment and control groups are removed.

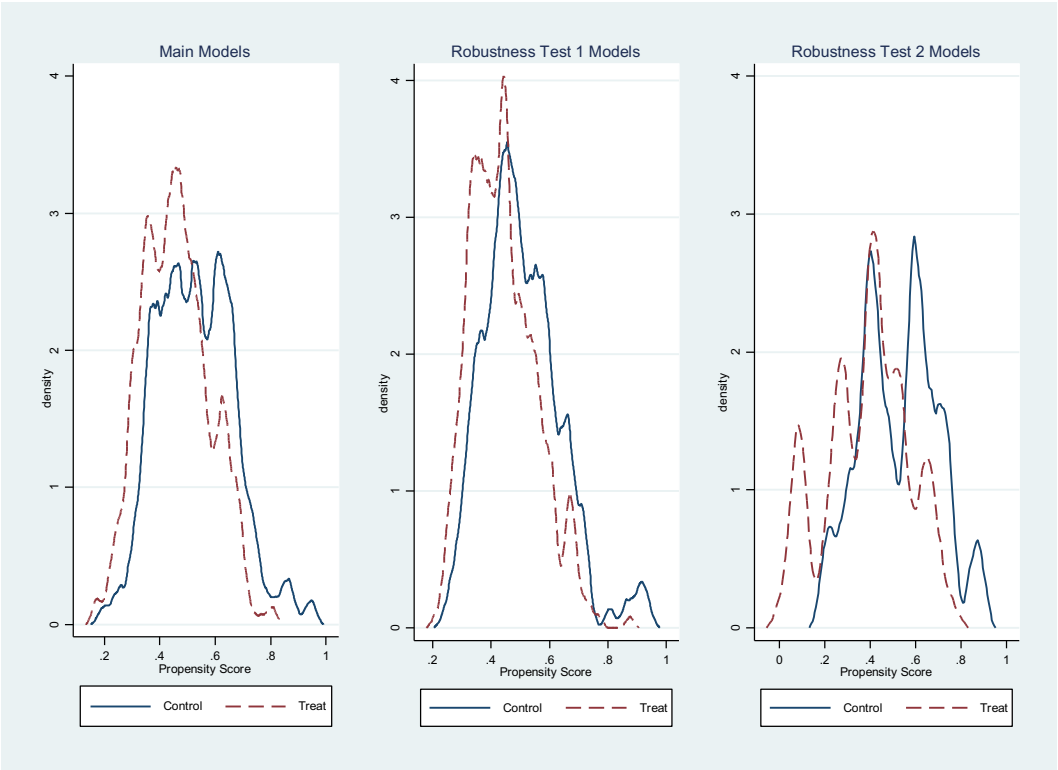
Moreover, findings in Post Test-1 Standardized score differences reported in Table 2 in the main text above correspond to the second row in Table B2 here, which shows a statistically significant causal effect of KA[9] when standardized test scores at Post Test-1 are considered as the outcomes. Additionally, as seen in third row, this unconfoundedness test also finds statistically significant results for standardized test scores at Post Test-2. These findings further confirm the plausibility of the unconfoundedness assumption, and supports that there is no compelling reason that the distribution of Pre-Test scores differs by treatment group, conditional on the remaining pre-intervention variables.

B3. Overlap assumption

Figure B3 depicts the distribution of propensity scores for Treatment and Control group students. As required under the overlap assumption, the plots do not indicate excessive probability mass near the two extremes 0 or 1, and estimated densities have most of their respective masses in overlap region. Hence, there is no evidence of the overlap assumption being violated.

[9] Similarly, corresponding to findings for Post Test-2 standardized test scores reported in Table 3, the third-row results in Table B2 are not statistically significant.

Figure B3: Distribution of Propensity Scores for Treatment and Control Group Students



Source: Author's calculations based on field data.