International Journal of Educational Development

Online Tutoring Reduces by Half the Learning Loss Due to School Closures: Evidence from a Randomized Experiment in Kenya

--Manuscript Draft--

Manuscript Number:	EDEV-D-23-01514R1
Article Type:	Full Length Article
Keywords:	Online Tutoring, Field Experiment
Corresponding Author:	Matthieu Chemin, PhD McGill University Montreal, QC CANADA
First Author:	Matthieu Chemin, PhD
Order of Authors:	Matthieu Chemin, PhD
	Jeremy Schneider
Abstract:	We evaluate the effects of an online tutoring program that started in 2016 and continued during the pandemic despite the schools being closed for 9 months in Kenya. Using videoconferences, volunteer students from a Canadian university tutored grade 6 students (12 years old) in a rural school in Kenya, on the topics of Maths and English. We implement a randomized experiment to test the effects. We find no effect when the schools are open, but a large effect when the schools are closed (0.4 SD increase in exam scores in the treatment group versus control group). Since we have data from before the pandemic, we are able to quantify the learning loss due to COVID-19: 0.8 SD. We conclude that online tutoring compensates half of the learning loss.
Suggested Reviewers:	Ana Dammert, PhD Associate Professor, Carleton University ana.dammert@carleton.ca Ana Dammert published in World Development. She has implemented randomized experiments in developing countries and is therefore qualified to referee this paper.
	Gabrielle Vasey, PhD Assistant Professor, Concordia University gabrielle.vasey@concordia.ca Gabrielle Vasey has published in the American Economics Journal: Applied Economics on distance education with remote rural areas and is thus uniquely suited to referee this paper
Opposed Reviewers:	
Response to Reviewers:	

Online Tutoring Reduces by Half the Learning Loss Due to School Closures: Evidence from a Randomized Experiment in Kenya

By

Draft: September 13, 2024

We evaluate the effects of an online tutoring program that started in 2016 and continued during the pandemic despite the schools being closed for 9 months in Kenya. Using videoconferences, volunteer students from a Canadian university tutored grade 6 students (12 years old) in a rural school in Kenya, on the topics of Maths and English. We implement a randomized experiment to test the effects. We find no effect when the schools are open, but a large effect when the schools are closed (0.4 SD increase in exam scores in the treatment group versus control group). Since we have data from before the pandemic, we are able to quantify the learning loss due to COVID-19: 0.8 SD. We conclude that online tutoring compensates half of the learning loss.

Two thirds of children fail to achieve a minimum proficiency level in reading and mathematics in grade 2[1] despite the ambitions of Sustainable Development Goal 4 for "inclusive and equitable quality education and lifelong opportunities for all." Tutoring - defined in Nickow, Oreopoulos and Quan (2020) as one-on-one or small-group instructional programming by teachers, paraprofessionals, volunteers, or parents - might be a valuable option: it causally improves grades (see Nickow, Oreopoulos and Quan (2020) for a review of the experimental literature), it is the ultimate customization of learning and reduction in class size, it allows for more engagement, rapid feedback, human connection and mentoring, and it bypasses the systemic issues of education systems in developing countries. The problem is how to reach students in remote rural areas of low-income countries, as well as high costs and the limited local supply of tutors.[2]

In this paper, we explore the potential of online tutoring by volunteers to address these issues. The recent improvements in communication technologies have made it possible for a tutor to teach students even in remote rural areas of developing countries. Having volunteers teach online can both drive down costs and expand the set of tutors available. Importantly, it can continue even if schools shut down. Despite the simplicity of the idea, there is no evidence that this would work in a remote rural area context of a developing country, where the efficacy of the treatment might be negatively affected by the cultural divide between tutors and tutees.

In this paper, we implement a randomized experiment on online tutoring in remote rural areas of Kenya. The tutors are university student volunteers. They communicate through the internet on an electronic tablet with their tutees. The tutees are primary school students in rural Kenya, at the grade 6 level (12 years old). The tutoring subject was English for the years 2016 to 2018, and Maths for the years 2019 to 2020.

A unique feature of our program is that it started in 2016 and because of its online nature, continued uninterruptedly after March 2020 when the schools closed in Kenya for 9 months. The Kenyan Government took time to respond by providing lessons through TV, radio, and the internet. These programs were widely criticized for being inaccessible, difficult to follow, and not adapted to the level of students in rural remote areas, further aggravating inequalities.[3] In contrast, the tutoring continued uninterrupted. We are thus able to evaluate the effects of the same program at two different points in time, when the schools are open and when they are closed.

We find little effects of this online tutoring program when the schools are open, and a large effect when the schools are closed. When the schools are open, the English tutoring has a modest effect

[1] Data from world development indicators.

[2] For example, Romero, Chen and Magari (2021) finds that tutoring with local tutors does not improve grades in Kenya.

[3] See for example Patrinos, Vegas and Carter-Rau (2022); Olanrewaju et al. (2021); Ochieng and Ngware (2022); Malenya and Ohba (2023); Mabeya (2020).

on reading comprehension, and the Maths tutoring has no effect. The results are very different when the schools are closed: we find a large effect on grades in that time period (0.4 SD in Maths, the discipline taught at that time). Thus, online tutoring appears especially effective when no other schooling options are available. Our explanation is decreasing returns to hours of teaching. When the schools are open, the tutoring program (1 hour of Maths per week) comes after a full teaching load (3 hours of Maths per week). We find little effects there. When the schools are closed, the online tutoring is the only source of education (barring the official TV/radio program). We find a large effect at that time.

Online tutoring appears to be critical in the period of school closures due to COVID-19. We dig deeper into this result by first quantifying the learning loss due to school closures, a subject of intense academic and policy debates, with estimates ranging from 0 to 0.7 SD, the higher estimates being found in remote rural areas.[4] The fact that we collected data before and after the pandemic allows us to quantify the learning loss in our context. We compare the evolution in scores of the 2020 cohort to the 2019 one (in the control groups). We find a 0.8 SD reduction in education achievement test scores, on the high end of the estimates provided in the literature, which is consistent with the local context of a remote rural area of a developing country with few alternative online options available. We conclude that the online tutoring program compensates for (0.4/0.8=) half of the learning loss.

The final finding concerns aspirations. We document a large loss in aspirations when the schools are closed, especially aspirations to go to university. An explanation is that students know that their chances to go to university have been harmed. The online tutoring program does not compensate for this: there is no discernible effect on aspirations in the treatment group compared to the control group.

Overall, we thus conclude that the tutoring program compensates partially (half) for the learning loss on cognitive skills, but does not compensate for the negative effect on aspirations. These results are important for policy implications: while online tutoring holds some promise (at least for cognitive skills), it does not fully substitute for school time. These large learning losses estimated in this paper as well as the large decrease in aspirations must be factored in when deciding on future school closures.

Our paper contributes to a burgeoning literature on tutoring from developed countries (Nickow, Oreopoulos and Quan, 2020; Carlana and La Ferrara, 2021; Kraft et al., 2022). Our study provides the first randomized experiment in developing countries, where education systems have systemic

[4]Singh, Romero and Muralidharan (2022); Moscoviz and Evans (2022); Patrinos, Vegas and Carter-Rau (2022); Engzell, Frey and Verhagen (2021); Maldonado and De Witte (2020); Kuhfeld et al. (2020); Azevedo et al. (2020); Hevia et al. (2022)

issues and online tutoring has a high potential to reach underserved communities.

In a developing country context, our paper also contributes to a growing literature on ways to mitigate the learning loss of closing schools. Previous studies have found positive results of projects providing SMS and 5-10-minute phone calls in Botswana and Nepal (Angrist, Bergman and Matsheng, 2022; Radhakrishnan et al., 2021), 30-minute phone calls by teachers in Bangladesh (Beam, Mukherjee and Navarro-Sola, 2022), 30-minute phone tutoring sessions in Bangladesh (Hassan et al., 2022), but no effects of teacher-student 15-minute mini-tutoring sessions in Kenya (Schueler and Rodriguez-Segura, 2021) or from weekly phone tutorials from teachers in Sierra Leone (Crawfurd et al., 2022). The contribution of our paper is to study online video tutoring. Importantly, our experiment starts prior to the pandemic, in 2016. This allows us to study online tutoring when the schools are open, and also to quantify the learning loss due to the school closures using data from before. We find a large 0.8 SD learning loss. We are thus able to answer the question of how much of the learning loss is mitigated by online tutoring (our answer is half). The scalability of online tutoring depends critically on the supply of college students who are willing to volunteer their time as tutors. The objective of the current paper is more to provide evidence on the likely effects of online tutoring: we find very limited effects of online tutoring when the schools are open, and a larger effect when the schools are closed.

I. Intervention

The intervention consists in offering free tutoring to primary school students living in a rural community of Kenya (Kianyaga, three hours north of Nairobi).[5] The innovative part of the program is that it is conducted online: the tutoring is done entirely by Skype (and then Zoom). The tutors are Canadian university students who volunteered to become tutors for the program. Students receive one hour of tutoring per week. Tutors and tutees are paired randomly and stay together throughout the school term.

The tutoring was in English for the years 2016-2018 and in Maths for the years 2019-2020. For English tutoring, tutors are trained to undertake "ice-breaking" activities in the first tutoring session in order to establish a relationship and to gauge the English knowledge and learning level of tutees.[6] The tutor then follows the official English textbook and helps tutees with their homework, keeping in mind the actual learning level of the students. The tutors are told that there is no point attempting difficult exercises if the tutee lacks rudimentary skills. Instead, tutors are advised

[5] See https://elimu.lab.mcgill.ca/pamoja.html for a short video on the program and pictures of the area.

[6] Tutors introduce themselves, and follow a list of questions to ask their tutees (for example, what is your favorite sport/game, movie/TV show, subject at school?). The tutor then asks "what surrounds you?", prompting the tutee to describe the place where he/she is. The tutor also undertakes a "would you rather. . . ?" activity to encourage the tutee to talk about him/herself.

to first build fundamental skills. The tutors are provided a range of techniques to teach at the right level, such as going back to easier exercises, building their own exercises, not following the textbook if they have a better idea or if they think the textbook does not follow a logical order. The emphasis is placed on teaching one simple thing right rather than many complicated ones.

A typical tutoring session in English consists of several minutes of tutors and tutees catching up with each other, followed by the tutee reading the most recent chapter of their English textbook. During the session, the tutor follows along the reading and are encouraged to interject and help their tutee with words that they may find difficult to pronounce and are encouraged to answer the questions of tutees. At the end of each reading, there are questions that both tutor and tutee discuss and cover, to test the tutee's reading comprehension skills on the passage that was just read.

For the math tutoring, tutors are instructed to follow the material currently being taught in the students' math class. Tutors follow the same method of gauging the level of each student, going back in the textbook if they see the students struggling, with the objective to build foundational skills while still following the Kenyan curriculum and the current textbook. Tutors made an effort to engage all tutees in the sessions and all students offered the tutoring attended most of the sessions.

A crucial aspect of the program is that it continued after March 2020 when the schools closed. We deployed tablets in students' homes and offered the data costs to connect to the internet for the single hour of tutoring per week (0.24 USD per one hour session per child). Access to internet was given only for this single hour per week. The tutors continued the exact same tutoring they were providing before. The tutors and tutees made sure to find a calm area. No tutors reported significantly more disturbance than when the tutoring was done in the school. The tutoring sessions were conducted at the exact same time as they would have had the schools been opened.

At that same time, alternative options to schools were offered in Kenya. The Kenya Institute for Curriculum Development and UNICEF provided pre-primary and primary lessons, through TV, radio, and internet uploads. Students could access the official education extension material, available on the Kenya Education Cloud (KEC) (see https://kec.ac.ke/).

Qualitative evidence suggests that few children were able to access these education extension efforts. For children able to access them, the remote lessons moved too quickly for them, were not at the right level, and did not explain the material or solutions in a manner they found accessible. More generally in Kenya, these programs have been widely criticized (Ochieng and Ngware, 2022; Malenya and Ohba, 2023; Mabeya, 2020).

It is in this context that we suggest the possibility of tutoring as an alternative. Tutoring can alleviate the concerns raised above: tutoring can be personalized at the right level, and it can reach even the rural underserved communities. Yet there was no study demonstrating rigorously the effects of online tutoring. Our paper is the first to do so. The policy implication of our paper is that tutoring can work as an alternative, especially when schools are closed.

II. Data

We use administrative data on grades taken 9 times during the year (three per trimester) for grade 6 students, who are typically 12 years old.[7] We use the last grade in the year before as the baseline grade, to estimate baseline cognitive ability. We thus have one pre-treatment wave and 9 post-treatment waves (T=10).

This large number of repeated waves allows us to have a high statistical power in this study. McKenzie (2012) recommends going beyond the single baseline-single endline paradigm in randomized experiments to include more post-treatment waves, especially if there is low autocorrelation in the outcome studied. In our case, there is a 0.53 autocorrelation in the Maths grades.

The total sample size is 2,439 observations.[8] This sample has enough statistical power to identify a minimum detectable effect size of 2.5 percentage points in grades.[9]

Even though our study is statistically powered to detect this effect, the downside of a small N sample is external validity. On key metrics, our sample is representative of the rest of Kenya. Students score on average 41 percent in Math and 231 out of 500 on all fields.[10] These scores are very similar to national averages.[11]

We complement the administrative data on grades with a survey, collected 4 times per year.[12] When the schools were open, we collected the survey in the school. When the schools were closed,

[7]There was an exception made in 2019 and 2020 when the number of grade 6 students was slightly too low and few grade 5 students were entered in the study. The tests are designed at the sub county level (schools within the sub county sit for similar tests).

[8]With one pre-treatment wave and 9 post-treatment waves (T=10) and 299 unique student-year observations, a balanced panel would contain (N*T=) 2990 observations of students' grades. Our panel has fewer observations (2439) for three reasons. First, schools were closed during waves 2, 3, 4 in 2020 due to the pandemic. There was not enough time for the test in wave 7, which is also missing. Second, we were unable to trace grades for grade 5 students in the 8th post-treatment wave of 2019. Finally, there is attrition, with 13 grades missing for the years 2019-2020. We find no differential attrition between the treatment and control groups, as shown in Table F1. Additionally, we implement a test for attrition and find the same results, as described below.

[9]With a significance level of 5%, statistical power of 80%, equal size between treatment and control groups (149 observations each), standard deviation of 14, one pre-treatment wave and 9 post-treatment waves, autocorrelation of 0.53 in the math grade and an ancova method, the minimum detectable effect size is 2.5 percentage points.

[10]Other fields are: English, Swahili, Science, Social Studies, and Religious Studies.

[11]Oketch and Mutisya (2013) report that the proportion of schools scoring 250 marks and above between 2002 and 2011 is 42%. Moreover, disaggregated grades by fields of study are available for Isiolo, a county not far from Kirinyaga county where the study is situated, and the average Math grade is 48 percent, average total grade 241. data available at: https://africaopendata.org/dataset/kcpe-2020-performance-in-isiolo-county

[12]Baseline surveys are conducted at the start of every school year in January, with three follow-up surveys at the start and end of the second term in May and August, and an endline survey at the end of the school year in late October.

we collected the survey in students' home. This was slightly harder than staying on school grounds and waiting for students to come to school, which explains the slightly smaller sample for 2020 (with 37 missing observations). Thus, our total sample size is 1159 observations instead of the theoretical 1196 observations.

The descriptive statistics in these surveys are also very similar to national averages. In this study, students are 11.8 years old on average.[13] Within this sample, the proportion of females is 44 percent, once again in line with the Kenyan average of 48 percent.[14]

The communities where the program is implemented share common features with other rural communities in the Central province of Kenya in particular, and Kenya in general. For example, the averages of age, gender, and poverty levels are similar to those of other communities in the 2009 Kenya Population and Housing Census; the 2005 Kenya Integrated Household Budget Survey (KIHBS); and the 2008 Kenya Demographic and Health Survey (DHS) (as found in Chemin (2018)). The particular area was selected in 2007 for a study on the effects of access to electricity, a project which has not yet fully materialized. Therefore, this community was not selected for this particular project on online tutoring.

We develop our own measure of oral proficiency in English, explained in greater detail in Appendix A, using the internationally recognized "Common European Framework of Reference for Languages (CEFR)". Table 1 shows that the average oral proficiency score in the baseline of the control group is 3.10 (out of 6), which corresponds to level A2 (basic user) in the CEFR classification.

We also ask questions on cross-cultural communication, on a scale from 1 (least) to 5 (most) how comfortable they would be talking to someone from another country and how much they would worry about what to say to someone from another country (details in Appendix IV.E). This section allows us to track how the intervention affects the student's comfort speaking and interacting with non-Kikuyu individuals. For many of the students, these interactions were their first times meeting someone who comes from outside the local community. Table 1 shows that the average is 3.87 out of 5, this includes the entire sample with the effect of the treatment.

We then ask questions on computer proficiency, explained in detail in Appendix G. This section is designed to track how the intervention affects the student's computer and technology proficiency over time. For many of these students, this was their first times using a computer, as evidenced by the very low average over these five questions (2.08 out of 5) in Table 1.

We also include in our survey measures on aspirations, related to higher education, career, and

[13] 27% of the whole sample comes from grade 5 since we included few grade 5 students in 2019 and 2020 to increase the sample size, as explained above.

[14] p.291 of the 2021 Economic Survey available at: https://www.knbs.or.ke/wp-content/uploads/2021/09/Economic-Survey-2021.pdf

broader goals in life. We ask students whether they desire to go to university, their desired age to marry and number of kids, what future career they would like to pursue, and other similar questions. Since questions are on different scale, we standardize all the variables, calculate the unweighted average, and re-standardize on the baseline wave. The purpose for these questions is to see how students may be motivated to continue staying in school. For example, if a student says that they would like to marry at a later age, this could indicate that the student wants to carry on with higher education and a career first, similar to their response on how many kids they would like to have. Since we also ask students what their desired future career would be, we can see whether students want to take on jobs that are more human-capital intensive and require higher education, such as lawyers, doctors, nurses, or if they want to take on other vocations which may not require formal schooling such as army or police officers, performers or professional athletes. With the intervention, we expect treated students to want to take on more human-capital intensive careers.

We also include other psychometric tools on liking school from Pell and Jarvis (2001), academic motivations from Muris (2001), self-esteem from Rosenberg et al. (1995), and perceptions of life in Canada and in Kenya to test whether the treatment affects these factors. All of the questions are explained in detail in Appendix G. In Appendix H, we find that the psychometric scales used in this paper display internal reliability, test-retest reliability, convergent validity, divergent validity and predictive validity.

Table 1 shows that the average response on the liking school index is 3.88 out of 5, motivation is 3.25 out of 5, self-esteem is 2.95 out of 5, and perceptions about Canada and Kenya is 0.93 and 0.86 out of 1 (where a value closer to 1 indicates a better perception). Overall, student generally like school, are motivated, and have a good perception of both Canada and Kenya.

III. Experimental Design

The way we randomized our sample is the following: we had a target number of 25 tutors per semester. We randomized half of the grade 6 students at the individual level into the treatment group. When the total size of the grade 6 class was more than 50 students, we simply selected 25 students from grade 6 to become the treatment group. When the total size was less than 50 students (this happened in the years 2019 and 2020 during the Math treatment), we randomized half of the class into the treatment group. This means less than 25 students are treated. Since our number of tutors was 25, we then consider grade 5 students and pick the rest of the treated students from grade 5. There is thus a treatment group and control group of grade 6 students, and

Table 1—Descriptive Statistics

	(1)	(2)	(3)
	Mean	SD	Count
Administrative data on test scores:			
Maths	41.3	13.89	2439
Grade Total	231.4	45.99	2435
Surveys:			
Age	11.81	1.16	286
School Year 5	0.27	0.44	290
Female	0.44	0.50	286
English Proficiency	3.06	1.19	1061
Cross-Culture Communication	3.86	0.79	1071
Computer Proficiency	2.08	0.99	1031
Aspirations	-0.27	1.00	1071
Liking School	3.88	0.35	1071
Motivation	3.25	0.52	1071
Self-Esteem	2.95	0.26	1071
Thoughts on Canada	0.93	0.16	1071
Thoughts on Kenya	0.86	0.12	1071

Note: Summary statistics for variables related to students' academic performances and baseline survey responses. Each of the variables after Female represent the baseline averages of an index consisting of various social and psychological questions related to the given topic. Apart from Aspirations, Thoughts on Canada, and Thoughts on Kenya, each of the indices can range from one to five. The aspirations index is standardized due to several of its components having different ranges, and the two indices related to thoughts on Canada and Kenya are comprised of variables that ranged from 0 to 1.

a treatment group and control group of grade 5 students.

When the treated students from grade 5 graduated to grade 6, we faced the choice of selecting them again for treatment in grade 6. This could have generated a treatment of 2 years for some. To keep things simple and limit the intervention to at most 1 year per student, we decided to exclude these treated students from the randomization of the next year. Thus, every treated student has at most received the treatment 1 year. We thus exclude these students treated when they graduate into grade 6, and select the new treated students from the rest of the sample.[15]

[15]A numerical example can be used here to illustrate the experimental design. Suppose first we have 60 students in grade 6. Our target number is 25 students treated, so we randomly draw 25 students to be treated, and 35 are control. In the final analysis, we then compare the 25 treated grade 6 students to the 35 control grade 6 students.

In another year, suppose we only have 30 students in grade 6. This can happen for reasons exogenous to the intervention, i.e., the cohort size shrinks in a particular year. We randomize half of grade 6 into treatment, such that 30/2=15 students are treated. Our target number is 25 students treated, such that 25-15=10 students still need to be treated. We thus consider grade 5 students. Suppose there are 40 students in grade 5. We randomly draw 10 students to be treated, and 30 are the control group. In the final analysis we compare the 15 treated grade 6 students to the 15 control grade 6 students, and the 10 treated grade 5 students to the 30 control grade 5 students, to control for the grade level. In practice, we implement this by having a dummy for grade 5 students, such that students of the same grade are compared with each other.

When the 40 grade 5 students graduate to grade 6, we excluded the 10 students already treated in grade 5 from the sample. Otherwise, some of these 10 students may have received 2 years of treatment, which would have complicated the analysis since we would then have to differentiate between one year of treatment and two years of treatment. We thus excluded these 10 students from the study, and only considered the 40-10=30 other students as part of the study. These 30 students were the control group in grade 5 and have thus not received any treatment. We then followed the same procedure, i.e., randomized half of that into treatment, compared the 15 treated to 15 control (excluding the 10 already treated in grade 5) and complemented

This randomization generates a variation in the number of treated students per classroom which is independent from the outcome studied, and only caused by our randomization process. In some classrooms, the number of treated students is 25. In others, the number of treated students is less, equal to half of the total size of the classroom. In yet other classrooms, the number of treated students is small, equal to the difference between the 25 tutors available and the number of students selected for treatment in grade 6.

We use these variations to identify peer effects. The basic idea of peer effects is that more treated students in a classroom should be associated with a positive effect on the control students. Importantly, the number of treated students in our case is independent from the outcome studied (the math grades) and are only related to the randomization process we used.

Aside from being able to measure peer effects, we argue that the experimental design sheds light on important questions. Recall that the schools closed in 2020. At that point, we distributed the tablets in the students' homes to continue the tutoring. We can compare the treatment effect in 2020 and in 2019, when the schools are closed or open. We are thus able to explore the temporal external validity of the results.

Moreover, we can quantify the learning loss due to the school closures by comparing the evolution of the control group in 2020 before and after schools resumed, compared to the evolution of the control group of the previous cohorts over the same time period.

We can then compare the treatment effect in 2020 to that learning loss to answer the question of how much of the learning loss is recovered by the program.

Table 2 shows the balance test. The important result from this table is that the grades are well balanced between the treatment and control group: the treatment group scores the exact same grade: 44 percent in Math and 195 on other fields of study. The differences are not statistically significant. The treatment group and control groups are thus comparable before the intervention starts.

The average age of the control group is 12 years old, 11.6 for the treatment group. There is a slight imbalance here. It is unclear whether older or younger students should react more or less to the treatment. We control for age in all regressions and find very similar results with or without this control.

Aside from this lone difference, none of the other variables are significantly different between the treatment and control groups.[16]

with the grade 5 students to reach our target number of 25 students treated.

[16]We obtained ethical approval for this study (REB File: 211-1015). There is no pre-analysis plan for this project designed in 2015, however we present in this paper all the outcomes of our questionnaire. We follow the recommendations of Banerjee et al. (2020), and present in the appendix the equivalent of a "populated" PAP, i.e., all the outcomes from the questionnaire. In this

Table 2—Balance Test: Treatment vs Control Group for Grade 6

	(1)	(2)	(3)	(4)
	Control	Treatment	Control-Treatment	P-value
Math Grade	44.21	44.38	-0.17	(0.94)
Grade Total (No Maths)	194.99	195.17	-0.18	(0.97)
Age	12.08	11.58	0.50*	(0.07)
Gender	0.42	0.49	-0.07	(0.33)
Other Cognitive Skills				
Oral Comprehension	2.91	2.71	0.19	(0.29)
Computer Proficiency	1.40	1.38	0.02	(0.87)
Cross-Culture Communication	3.48	3.46	0.02	(0.86)
Non-Cognitive Skills				
Aspirations	0.16	-0.02	0.18	(0.21)
Liking School	3.83	3.80	0.04	(0.34)
Motivation	3.15	3.12	0.04	(0.60)
Self-Esteem	2.94	2.89	0.05	(0.15)
Thoughts on Canada	0.98	0.96	0.02	(0.39)
Thoughts on Kenya	0.88	0.86	0.01	(0.44)

Note: Two-sample t-test results for baseline averages of variables related to students' academic performances and survey responses between treatment and control group. Columns 1 and 2 show the mean of the variable at baseline for the control and treatment groups respectively. Column 3 reports the t-test for the equality of means in the control and treatment groups, and column 4 shows the p-value of that difference. The baseline grades for Maths and Grade Total are taken as the final grades from wave 9 of the previous year.

IV. Empirical Analysis

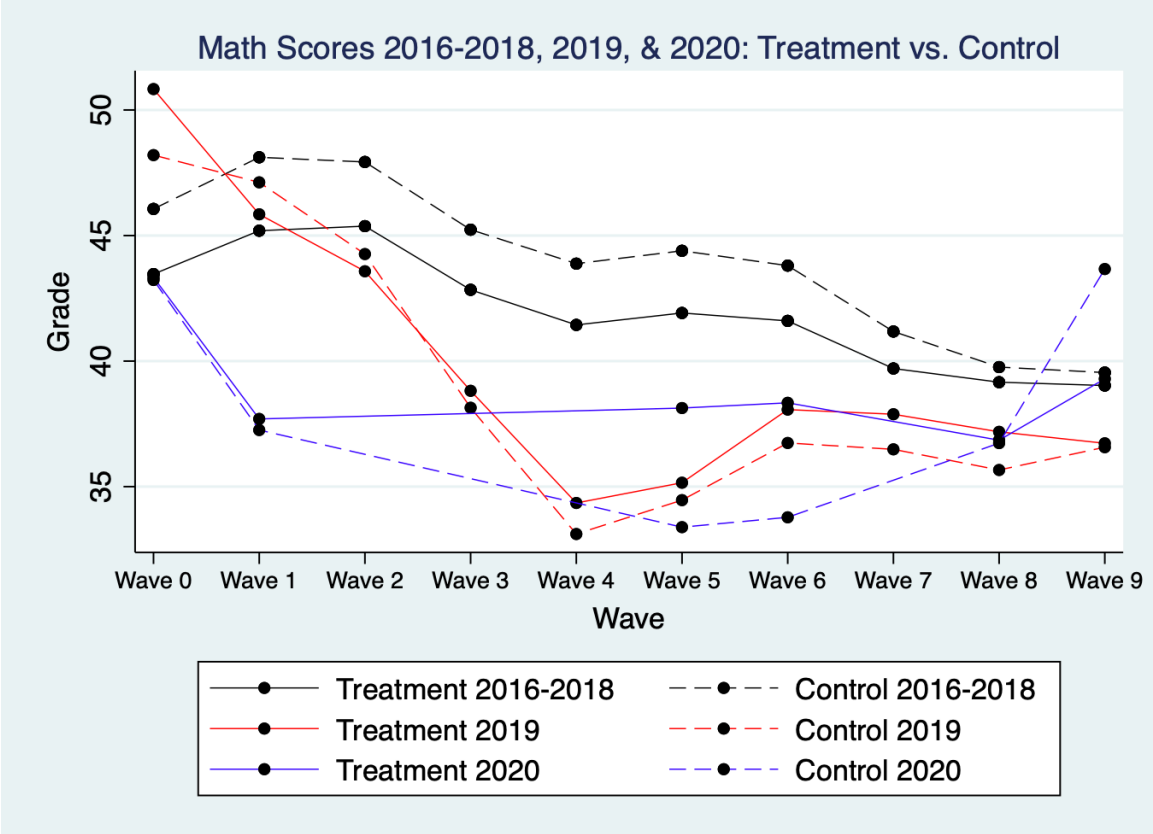
A. Effects on Math Grades

We show the raw data on Math grades in Figure 1 below. Wave 0 is the baseline and the treatment is implemented for waves 1 through 9. The black lines show the 2016-2018 period when the intervention was in English. The treatment group is in a solid line, and the control group is in a dashed line. As can be seen on the graph, the treatment has no effect on Math grades, which is logical since the intervention was in English at that time. This is actually reassuring for the integrity of the experiment: the treatment group and control group are on very similar trends absent the treatment (in Maths).

For the year 2019 (in red), we also see no effect of the Math tutoring program. Recall that the schools were open at that time.

paper, we depart from presenting all these outcomes as in a populated "PAP" since we made an important ex-post discovery: we found no effect of the intervention when the schools were open and an effect when the schools were closed. This allows us to estimate a production function of grades featuring decreasing returns, which we use to simulate the effect of closing schools. We had not pre-specified this approach since there was no way of knowing ex-ante that the pandemic would close down the schools for 9 months in March 2020. This new exposition of the results is in line with (Banerjee et al., 2020)'s recommendation of "presenting in the paper what was actually learned in the course of the experiment, as opposed to what was anticipated ex-ante".

Figure 1. Math Grades: 2016-2018 vs 2020



Note: The figure shows the raw math grades. Wave 0 represents the baseline Math grades (calculated as the last grade of the student in the year before). Waves 1-9 represent the respective periods in the school year. The Kenyan school year begins in January and is divided up into three trimesters, with each trimester containing three periods (and thus nine total periods in a school year). The period 2016-2018 is in back (English tutoring). The treatment group is the solid line, the dashed line is the control group. The year 2019 is in red (Maths tutoring, schools open). The year 2020 is in blue (Maths tutoring, schools closed). Schools were closed for waves 2 through 5 in 2020, hence the missing grades, but online tutoring continued.

The main result comes from 2020 (in blue). There is a large noticeable drop in math scores in the 2020 year between wave 0 and wave 1; however the drop is similar in the treatment group and control group. This large drop in both groups may be coming from the varying difficulty of exams. This highlights the importance of having a control group, to control for the difficulty of the exam.

A difference between the treatment and control groups emerges in later waves. Schools were closed for waves 2 through 4 in 2020, hence the missing grades, but online tutoring continued. In wave 5 when the schools reopen and grades are taken again, the treatment group is above the control group by a noticeable 5 percentage points difference.

The effect disappears over time as the schools reopen. In fact, the control group seems to outperform the treatment group by wave 9, although this effect is not statistically significant (see Figure B1 in Appendix B with confidence intervals). The only significant result in this graph is the positive effect of the treatment when schools were closed.

Thus, we conclude that the math tutoring program has little effect when the schools are open, but has a noticeable impact when the schools are closed. An explanation is that during the pandemic, this online tutoring program was the only source of education (barring the official TV/radio program). Thus, one hour of math tutoring has a large impact when the schools are closed, much less so when the schools are open.

To gauge whether this finding is statistically significant, we use the following specification:

$$y_{itk} = \beta_1 Math_{it} + \beta_2 SchoolClosed_{itk} + \beta_3 Math * SchoolClosed_{itk} + \beta_4 English_{it} + \beta_5 BaselineMathGrade_{it0} + \beta_6 BaselineMissing_{it0} + \beta_7 X_{it} + \beta_8 \delta_{tk} + \varepsilon_{itk} \quad (1)$$

where y_{itk} is the dependent variable in year t, wave k for student i, $Math_{it}$ is a dummy variable equal to 1 if the student was in the Math treatment group (in the years 2019 and 2020), $SchoolClosed_{itk}$ is a dummy variable equal to 1 if year $t = 2020$ and wave $k = 5$, and $Math * SchoolClosed_{itk}$ is the interaction of the Math Treatment dummy and the School Closed dummy. $English_{it}$ is a dummy variable equal to 1 if student i was in the treatment group (in the years 2016 to 2018).

The regression includes the full set of wave $\times$ year fixed effects δ_{tk}. Keeping in line with Figure 1, we aggregate the years 2016-2018 together in the wave fixed effects. For example, there is one dummy for wave 5 $\times$ Years 2016-2018. We report this coefficient in the main table, to show that there is nothing special about wave 5 in other years. The results are exactly the same if we disaggregate the years 2016 to 2018 in different wave*year fixed effect. Wave fixed effects for the years 2019 and 2020, however, remain separate, since these years are different due to the pandemic. In each column, the wave 1 of 2016-2018 is the omitted wave in the regression.

Some students are missing baseline grades in a given year, but have grades available throughout the school year period. To avoid losing this data in the regressions we run, we use the following method: for students that have baseline values available, this value is represented in the control variable $BaselineMathGrade_{it0}$. If, however, the baseline value is not available, the value of the control variable $BaselineMathGrade_{it0}$ is set to zero and a dummy variable $BaselineMissing_{it}$ is set equal to one. This allows us to keep all of the data available even when the baseline value is missing.

X_{it} represents a vector of control variables, including age, gender, school year, and students' baseline responses to survey questions related to their levels of motivation, self-esteem, future aspirations, how much they like school in general, and how much they enjoy their classes. Table 3 shows the results of this regression for students' math grades in the school years 2016-2020.

Standard errors are clustered at the student level.

Column (1) of Table 3 shows that the variable $Math_{it}$ is not statistically different from 0, indicating that being in the educational program did not have any significant effects, at least when the schools are open. The result changes when the schools are closed: the coefficient of $Math * SchoolClosed_{it}$ is statistically significant at the 5% level and has a coefficient of 5.67, indicating that in wave 5 of 2020, being in the treatment group was associated with a math score 5.67 points higher compared to the control group, exactly like in Figure 1.

The variable *SchoolClosed* (which is simply the dummy for wave 5 of 2020) measures the learning loss, according to the existing difference-in-differences literature, by comparing the evolution of the control group of the 2020 cohort to the control group of the 2016-2018 cohorts. This estimate of learning loss has to be interpreted with caution since it relies on the (untestable) parallel trends assumption, i.e., the 2020 cohort would have evolved the same way as the 2016-2018 cohorts absent the pandemic. Results indicate a decrease in math scores by 12 points. If we interpret this as the learning loss due to the school closures, then this would mean that the educational program alleviates ((5.67/11.95)*100)=47% of the learning loss. This is exactly what is shown in Figure 1, where the difference in math scores between the control group of 2016-2018 and the control group of 2020 is nearly 12 points between wave 5 and wave 0, and only 7 points in the treatment group.

An interesting check of the difference-in-differences approach is provided by the variable "Wave 5 * 2016-2018" (to reiterate, the full set of wave × year fixed effects is included, we choose to report only this coefficient in the table because it provides an interesting check). It is not significantly different from zero, which shows that there are no differences between wave 5 and wave 1 in the years 2016-2018. Thus, if we are willing to make the assumption that there would not be any difference either between wave 1 and wave 5 in the 2020 cohort, then the coefficient of *SchoolClosed* can be interpreted as the learning loss.

The results remain stable when we control for several variables in the rest of the table. Column (2) adds baseline grades excluding Math as a control variable to the specification in Column (1). The coefficients are largely unchanged. Therefore, controlling for baseline ability doesn't affect the results. Column (3) further builds on the previous specification by additionally controlling for various student characteristics included in the variable (i.e., age, gender, and the school year that the student is completing).

We also include the baseline value of the indices of each of the 10 sections of our survey, namely English oral comprehension, computer proficiency, cross-culture communication, motivation, self-esteem, aspiration, liking school, liking courses, thoughts on Canada, and thoughts on Kenya. The results are exactly the same when we control one by one for each of these indices as in Table

Table 3—Math Grades: 2016-2018 vs 2020

	(1)	(2)	(3)	(4)
	Dependent Variable: Math Grade			
Math	-0.66	-0.68	-0.52	-0.54
	(1.20)	(1.13)	(1.14)	(1.18)
Math * School Closed	5.80**	6.46**	6.61**	6.33**
	(2.69)	(2.66)	(2.71)	(2.76)
Fisher (p-val)	(0.055)	(0.029)	(0.038)	(0.051)
Wild Cluster Bootstrap (p-val)	(0.11)	(0.11)	(.1)	(0.075)
Attrition: Lower Bound	4.82*	5.29*	5.28*	5.29*
	(2.78)	(2.75)	(2.82)	(2.89)
Attrition: Upper Bound	6.38**	6.92**	7.01**	7.15**
	(2.79)	(2.74)	(2.75)	(2.81)
School Closed	-11.95***	-11.00***	-11.55***	-11.68***
	(2.09)	(2.19)	(2.40)	(3.56)
Wave 5 * 2016-2018	-0.22	-0.11	-0.10	-0.16
	(1.13)	(1.13)	(1.13)	(1.13)
English	-1.09	-0.81	-1.00	-1.06
	(1.39)	(1.32)	(1.30)	(1.20)
Wave*Year fixed effects	YES	YES	YES	YES
Controls:				
Baseline Grade	NO	YES	YES	YES
Age, Gender, School Year	NO	NO	YES	YES
Baseline Survey	NO	NO	NO	YES
Observations	2,170	2,170	2,170	2,170
R-squared	0.355	0.393	0.399	0.431
Mean Dep. Var	40.82	40.82	40.82	40.82
SD	13.77	13.77	13.77	13.77

Note: Standard errors clustered at the student level in parentheses. *** p<0.01, ** p<0.05, * p<0.1. For each column, the dependent variable is a student's math grade in a given wave. 'Math' is a dummy equal to 1 if a student was in the Math treatment group for the years 2019-2020. 'School Closed' is a dichotomous variable equal to 1 when schools are closed, i.e., wave 5 of 2020. 'MathTreatment * School Closed' is the interaction between the two variables. 'Wave 5 * 2016-2018' is a dummy equal to 1 if the period is wave 5 in any of the years 2016 - 2018. 'English' is a dummy variable equal to 1 if a student was in the treatment group for the years 2016 - 2018. All regressions include a full set of interactions between the 9 waves and the 3 time periods (2016-2018 for the English treatment, 2019 for the Math treatment, and 2020 for the Math treatment combined with school closure). All regressions control for baseline math grade and a dummy variable 'Baseline Missing' that is equal to 1 if the baseline math grade is missing. If baseline math grade is missing, it is replaced by the value 0 and the dummy variable 'Baseline Missing' takes the value 1. Column 1 shows the estimation of Equation 1 without any additional controls. Column 2 augments this specification by including the baseline grade on all other topics than Maths, and a dummy variable 'Baseline Missing Total Grade' equal to 1 if the baseline total grade is missing. Column 3 adds to column 2 by controlling for a student's age, gender, and the year of schooling they are currently completing. Finally, column 4 includes in the list of controls the baseline averages of various indices from the survey data. These indices include: oral comprehension, computer proficiency, cross-culture communication, motivation, self-esteem, future aspirations, liking school, liking classes, thoughts about Canada, and thoughts about Kenya.

C1 in Appendix C; or all of them together in Column (4) of Table 3. We also control for the 51 individual components of these indices, one by one or together, and still find the same effect for $Math * SchoolClosed_{itk}$ as shown in Table C2. Our results are thus not driven by baseline differences in cognitive or non-cognitive skills between the treatment and control groups.

The results are also the same if we disaggregate the wave fixed effects for the years 2016 to 2018 into different dummies, as can be seen in Table D1 in Appendix D. The results are very similar if we restrict the sample to the years 2019 and 2020 alone when the tutoring was in Maths, as shown in Appendix E Table E1.[17]

B. Robustness Checks

We present three robustness checks to adjust the standard errors for the small number of students. First, we use the exact Fisher test (Young, 2018). This permutation test is an exact test regardless of sample size or distribution of error term, as opposed to conventional t-tests which depend on the assumption of large samples (to use asymptotic results), a condition that may be violated in the sample we use, or a normal distribution of the error term. To implement this procedure, we obtain the observed t-stat for the outcome in question, permute the observations randomly between the treatment and control groups, obtain a simulated t-test, repeat this 1,000 times, and find the proportion of occurrences the simulated t-stat is above the observed t-stat, which is the Fisher p-value. In Column (1), the Fisher p-value is 0.055.

Second, we provide a test for the clustering. In our preferred specification, we cluster the standard errors at the level of students. Yet, they could also be clustered at the level of cohorts, which are few (6 cohorts during 4 years). We use the Wild Cluster Bootstrap methodology described in Cameron, Gelbach and Miller (2008) to address this issue. Using Monte Carlo simulations with 6 clusters and different error structures and cluster sizes, they show that cluster-robust standard errors reject the null at a rate of 8.2 percent to 18.3 percent. The intuition of the Wild Cluster Bootstrap methodology is to resample residuals at the level of a cluster, thereby preserving the clustering of the data. With 6 clusters, they show that this technique rejects the null at a rate of 1.9 percent to 5.3 percent, not significantly different from 5 percent. In our analysis, we use the 6-point weight distribution proposed by Webb (2014). We find that the results are robust to this correction, especially in the most preferred specification when adding controls in Columns (3) and (4).

Finally, we address the issue of attrition. There is little attrition in the math grades since the

[17]The results are also the same if we look at the value-added compared to wave 0 in our specification.

data is administrative at the school level (13 missing observations for the years 2019-2020). We find that there is no differential attrition between the treatment and control groups, as shown in Table F1. Moreover, we propose a test for attrition using Manski bounds. We replace the missing observations in the treatment group by the minimum observed value, and in the control group by the maximum value. This represents in a way a worst-case scenario for our estimate. Column (1) of Table 3 shows that the main result is still statistically significant in this worst-case scenario. We also present the best-case scenario, in which we replace the missing observations in the treatment group by the maximum observed value, and in the control group by the minimum value. This builds an upper bound for our estimate. Since the lower bound of the worst-case scenario is still statistically significant, we conclude that the problem of attrition is unlikely to bias our estimates.

C. Peer Effects

Recall that because of the way we randomized, there is exogenous variation in treatment intensity: in some classrooms, there were 25 students treated, in others less (if the class size was below 50 students) and in grade 5 classrooms, there were few students treated (to be precise, the difference between 25 and the number of students treated in the grade 6 classroom since our target number of tutors was 25). These variations are exogenous to the outcome studied and solely dependent on our randomization process.

We simply count the number of students treated by the Math intervention per classroom in a variable called "Number Treated Math" and include it in our regressions. More treated students should be associated with a better performance of the control group according to the logic of peer effects. Since the variable "Math" is included, this variable must be interpreted at Math=0, i.e., it represents the increase in Math grades in the control group due to a greater number of students treated by the Math intervention.

Column (1) of Table 4 repeats the main analysis. Column (2) of Table 4 adds this new variable "Number Treated Math". We find no effect of this variable on the Math grades. In fact, the inclusion of this variable makes no difference to the main coefficient of "Math * School Closed" studied in this paper. More treated students do not lead to a better performance of the control group.

The results are the same in Column (3) if we consider the proportion of students receiving the intervention rather than the number, to account for different class sizes.

The failure of our statistical test in Table 4 to detect peer effects and the natural absence of peer effects when the schools are closed make it unlikely that our results would be confounded by any

Table 4—Peer Effects in the Classroom

	(1)	(2)	(3)
	Dependent Variable: Math Grade		
Math	-0.61	-2.36	0.69
	(1.14)	(2.79)	(2.62)
Math * School Closed	6.33**	6.08**	6.38**
	(2.67)	(2.67)	(2.70)
School Closed	-11.24***	-11.24***	-11.24***
	(2.13)	(2.14)	(2.13)
Wave 5 * 2016-2018	-0.10	-0.10	-0.10
	(1.13)	(1.13)	(1.13)
English	-0.88	-0.88	-0.88
	(1.31)	(1.31)	(1.31)
Number Treated Math		0.14	
		(0.22)	
Proportion treated Math			-3.24
			(6.46)
Observations	2,170	2,170	2,170
R-squared	0.392	0.393	0.392
Mean Dep Var	40.82	40.82	40.82
SD	13.77	13.77	13.77

Note: Standard errors clustered at the student level in parentheses. *** p<0.01, ** p<0.05, * p<0.1. For each column, the dependent variable is a student's math grade in a given wave. 'Math' is a dummy equal to 1 if a student was in the Math treatment group for the years 2019-2020. 'School Closed' is a dichotomous variable equal to 1 when schools are closed, i.e., wave 5 of 2020. 'MathTreatment * School Closed' is the interaction between the two variables. 'Wave 5 * 2016-2018' is a dummy equal to 1 if the period is wave 5 in any of the years 2016 - 2018. 'English' is a dummy variable equal to 1 if a student was in the treatment group for the years 2016 - 2018. All regressions include a full set of interactions between the 9 waves and the 3 time periods (2016-2018 for the English treatment, 2019 for the Math treatment, and 2020 for the Math treatment combined with school closure). All regressions control for baseline math grade and a dummy variable 'Baseline Missing' that is equal to 1 if the baseline math grade is missing. If baseline math grade is missing, it is replaced by the value 0 and the dummy variable 'Baseline Missing' takes the value 1. In Column (2), the variable "Number Treated Math" is the number of students treated by the Math intervention per classroom. In Column (3), the variable "Proportion Treated Math" is the proportion of students treated by the Math intervention in the classroom.

peer effects.

D. Effects on English Proficiency

The results above focus on the effect of the Math intervention on the Math grades. We now turn to the English intervention. The first finding from Table 3 relates to the coefficient *English*: the online tutoring intervention in English (organized in 2016-2018) did not increase the math grades. One could have expected a positive impact there since the math textbook is in English, whereas students' home language in this area is Kikuyu, the local dialect. A better mastery of English could have increased the math grades. This is not what we find.

To look more directly at the effects of the English intervention on English proficiency, we use our own assessment tool of oral comprehension in English explained in detail in Appendix A and that follows the in the CEFR classification. This information was collected in our surveys collected 4 times a year (hence the smaller sample size). We build an index of four measures: understanding, conversation, vocabulary, and spoken fluency.

Column (1) of Table 5 shows that the average oral proficiency score in the baseline of the control group is 3.07 (out of 6), which corresponds to level A2 (basic user) in the CEFR classification.

The English tutoring (implemented in the 2016-2018 period) increases this outcome by a statistically significant 0.21 (out of 6). This corresponds to a (0.21/1.215 =) 0.17 standard deviations increase in overall oral comprehension. Thus, online tutoring in English is associated with beneficial effects on English proficiency.

We also find that school closures had a detrimental effect on oral comprehension. Just in the period of school closures alone, the average oral comprehension level dropped by more than 1 standard deviation (1.43/1.215). The effect of the school closure was not compensated by the Math tutoring intervention, which is quite logical since the tutoring was in Math.

E. Effects on Cross-Cultural Communication

A key question with online tutoring in a cross-cultural context is whether the cultural divide may negatively affect the tutoring.

In fact, we find in Table 6 shows that the treatment leads to overall higher student capabilities in cross-cultural communication. Column (1) shows the unweighted average of our two questions on the topic: "How comfortable would you be talking to somebody from another country?", and "How much would you worry about what to say if you were talking to someone from another country?".

The tutoring in English improves cross-cultural communication comfort by (0.41/0.791) = 0.52

Table 5—Oral Comprehension

	(1)	(2)	(3)	(4)	(5)
	Index	Understanding	Conversation	Vocabulary	Spoken Fluency
Math	0.18	0.15	0.11	0.23*	0.19
	(0.12)	(0.14)	(0.15)	(0.12)	(0.14)
Math * School Closed	-0.38**	-0.33	-0.29	-0.22	-0.25
	(0.17)	(0.21)	(0.21)	(0.20)	(0.23)
School Closed	-1.43***	-1.17***	-1.41***	-0.84***	-1.09***
	(0.24)	(0.26)	(0.24)	(0.24)	(0.25)
English	0.21*	0.21	0.22	0.14	0.20
	(0.13)	(0.14)	(0.13)	(0.14)	(0.14)
Observations	812	813	813	813	812
R-squared	0.491	0.435	0.434	0.370	0.441
Mean Dep. Var	3.074	3.381	3.043	2.943	2.927
SD	1.215	1.254	1.292	1.204	1.351

Note: Standard errors clustered at the student level in parentheses. *** p<0.01, ** p<0.05, * p<0.1. In Column 1, the dependent variable is the unweighted average of the four components. Columns 2 through 5 show the results of the same regression specification but with each individual component of the oral comprehension index as the dependent variable. The components of this index all range from 1 to 5, with 1 indicating a poor oral comprehension and 5 indicating strong comprehension, and include: understanding, conversation, vocabulary, and spoke fluency. All regressions control for a student's age, gender, and current year of schooling, as well as the baseline survey response of the dependent variable and a dummy variable "Baseline Missing" equal to 1 if the baseline survey response is missing. If baseline survey response is missing, it is replaced by the value 0 and the dummy variable "Baseline Missing" takes the value 1.

standard deviations compared to the control group. Similarly, students who received tutoring in Math reported being more comfortable in cross-cultural communication, although the effect was less pronounced (0.16/0.791 = 0.20 standard deviations). This is logical since there may be less interactions in the Math tutoring than in the English tutoring.

Table 6—Cross-Culture Communication

	(1)	(2)	(3)
	Index	Talking to someone from other country	Inverse: Worry when Talking to someone from other country
Math	0.16*	0.16**	0.13
	(0.08)	(0.08)	(0.11)
Math * School Closed	0.04	0.12	0.01
	(0.15)	(0.16)	(0.17)
School Closed	0.05	0.03	0.19
	(0.14)	(0.15)	(0.18)
English	0.41***	0.41***	0.40***
	(0.09)	(0.08)	(0.11)
Observations	821	822	820
R-squared	0.212	0.155	0.239
Mean Dep. Var.	3.922	3.943	3.900
SD	0.791	0.746	1.065

Note: Standard errors clustered at the student level in parentheses.. *** p<0.01, ** p<0.05, * p<0.1. Column 1 shows the estimation of equation 1 with the unweighted average of the cross-cultural communication index as the dependent variable. Columns 2 and 3 show the results of the same regression specification but with each individual component of the cross-cultural communication index as the dependent variable. The components of this index all range from 1 to 5, with 1 indicating the least amount of comfort and 5 indicating the highest level of comfort. They include: talking to someone from another country, and worrying about what to say when talking to someone from another country. Because column 3 asks a question where the ideal response is the lowest possible value, we reverse the response values (i.e. a response of 5 out of 5 for the question in column 3 now indicates that a student doesn't worry at all when talking to someone from another country). All regressions control for a student's age, gender, and current year of schooling, as well as the baseline survey response of the dependent variable and a dummy variable "Baseline Missing" equal to 1 if the baseline survey response is missing. If baseline survey response is missing, it is replaced by the value 0 and the dummy variable "Baseline Missing" takes the value 1.

In Appendix A, we also find a strong effect of both interventions (Math or English) on computer skills in Table G1. This is logical since for some students, this was the first time they were using a tablet with an internet connection.

F. Effects on Aspirations

We now turn to aspirations. Table 7 follows the same empirical specification, with the dependent variable in column 1 reflecting the standardized average of all questions in the Aspirations questionnaire and the remaining columns displaying the results for each individual component of the

index.

Column 1 shows that the period of school closures is associated with a lower average Aspiration index score by 1.06 standard deviations. The coefficient of $Math{*}SchoolClosed_{it}$ is not significantly different from zero, indicating that the online tutoring program does not compensate for this loss in aspirations.

Table 7—Aspirations

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Index	Likely university	Desired # of kids	Spend on education?	Best job?	Desired job?	Motivated?	# of hours to study
Math	0.15	0.14	-0.18*	-0.03	0.30**	0.33**	-0.02	0.12
	(0.10)	(0.09)	(0.10)	(0.10)	(0.15)	(0.15)	(0.15)	(0.11)
Math * School Closed	0.29	0.22	0.49*	-0.31	0.44	0.28	0.03	-0.02
	(0.25)	(0.21)	(0.27)	(0.22)	(0.39)	(0.37)	(0.15)	(0.15)
School Closed	-1.06***	-0.36*	-0.25	-0.93***	-0.80***	-0.60**	0.11	-0.61***
	(0.18)	(0.18)	(0.17)	(0.19)	(0.28)	(0.27)	(0.13)	(0.13)
English	-0.18**	-0.08	-0.08	-0.09	-0.16*	-0.15*	-0.04	-0.02
	(0.08)	(0.09)	(0.07)	(0.08)	(0.09)	(0.09)	(0.11)	(0.09)
Observations	822	822	777	823	744	722	820	818
R-squared	0.244	0.203	0.177	0.303	0.201	0.199	0.204	0.230
Mean Dep. Var	-0.362	-0.280	-0.171	-0.0455	-0.105	-0.107	-0.159	-0.321
SD	0.986	0.956	0.853	1.001	1.117	1.116	1.281	0.984

Note: Standard errors clustered at the student level in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Column (1) shows the estimation of equation 1 with the aspirations index as the dependent variable. In Column (2), the question is "How likely are you to go to university?" on a scale from 1 (definitely not go) to 5 (definitely will go. In Column (3), the dependent variable is the desired number of kids (inverted since we interpret a high response as having low aspirations; and standardized). In Column (4), the question is "If you were given 1000 Kenyan Shillings, how would you spend it?". Answers which related to school expenditures (e.g. bags, textbooks, uniforms, pens, pencils) were coded as 1 and other non-school related expenditures (e.g. toys, cell phone, radio, TV) were coded as 0. In Column (5), the question is: "What do you think is the best job in the world?". Answers which typically require higher education (e.g. doctor, nurse, engineer, lawyer) were tagged as 1 and other occupations (e.g. police man, soldier, football player) were tagged as 0. In Column (6), the question is: "Do you know what job you want to have in the future?". We re-code responses to this question in the same manner as above. In Column (7), the question is: "On a scale from 1 (not at all) to 10 (extremely motivated), how motivated are you to work hard?". In Column (8), the question is "How many hours per day would you be willing to spend on school work in order to go to university?". A high response to these previous two questions indicates high student aspirations. We standardize this variable. All variables in columns (2) to (8) are standardized, added together in an unweighted average, and re-standardized to make up the aspirations index of Column (1). All regressions control for a student's age, gender, and current year of schooling, as well as the baseline survey response of the dependent variable and a dummy variable "Baseline Missing" equal to 1 if the baseline survey response is missing. If baseline survey response is missing, it is replaced by the value 0 and the dummy variable "Baseline Missing" takes the value 1.

The rest of the table shows that this reduction in aspirations during the school closures comes from a decrease in aspirations to go to university (column 2), a reduction in the willingness to invest in one's education (column 4), a reduction in viewing high-skilled jobs as the best job or even a desirable job in the future (columns 5 and 6), and a reduction in the number of hours of work per

day that they are willing to spend in order to go to university (column 8). School closures thus had a negative effect on aspirations.

The coefficient on the interaction of school closures and the treatment dummy isn't statistically different from 0. Thus, the tutoring program didn't salvage the lost aspirations experienced by the students due to the school closures.

In Appendix G, we present other indices such as motivation in Table G4, self-esteem in Table G5, and perceptions about Canada in Table G6 or Kenya in Table G7 and find very little effects of either interventions.

G. Discussion

The main result of the paper is that the online tutoring increases grades in Math when the schools are closed, but not when they are open. One explanation for this finding is decreasing returns to education.

In Appendix I, we use this fact to propose a methodology to estimate the learning loss, other than relying on difference-in-differences. We fit a model with decreasing returns to hours of math studied, using the exogenous variation provided by the randomized experiment implemented at two different points in time, after 0 hours studied (when the schools are closed) and 3 hours studied (when the schools are open). After estimating the model, we then use it to simulate school closures (i.e., going from 3 to 0 hours studied).

We find an estimate very close to the difference-in-difference estimator, and which does not rely on the parallel trends assumption. Instead, our estimator relies on a randomized experiment, implemented at two different points in time, such that we can evaluate the decreasing returns to hours of teaching in math in a production function of grades. The fact that these two methodologies yield relatively similar estimates support the claim that school closures causally created a large learning loss. This is important because most of the literature on quantifying the learning loss has been relying on a difference-in-differences estimate, which appears to be a valid estimator for the learning loss in our context.

V. Conclusion

School closures at the beginning of the COVID-19 pandemic had profound impacts on students' learning across the world. Governments around the world tried to put measures in place to address the learning loss. For example, in Kenya, the government introduced online distance learning initiatives through TV, radio, and internet uploads. These programs have been widely criticized by the

literature for being inaccessible, especially in rural areas (Ochieng and Ngware, 2022; Malenya and Ohba, 2023; Mabeya, 2020). In this paper, we suggest the possibility of tutoring as an alternative. Tutoring can alleviate the concerns raised above: tutoring can be personalized at the right level, and it can reach even the rural underserved communities. Yet, no studies rigorously demonstrated the effects of online tutoring. Our paper is the first to do so. The policy implication of our paper is that tutoring can work as an alternative, especially when schools are closed.

Our study also adds to the literature about the effect of school closures on academic achievement. This has been the subject of an intense academic and policy debates, with estimates ranging from 0 to 0.7 SD, the higher estimates being found in remote rural areas (Singh, Romero and Muralidharan, 2022; Moscoviz and Evans, 2022; Patrinos, Vegas and Carter-Rau, 2022; Engzell, Frey and Verhagen, 2021; Maldonado and De Witte, 2020; Kuhfeld et al., 2020; Azevedo et al., 2020; Hevia et al., 2022). The fact that we collected data before and after the pandemic allows us to quantify the learning loss in our context. We compare the evolution in scores of the 2020 cohort to the 2019 one (in the control groups). We find a 0.8 SD reduction in education achievement test scores scores, on the high end of the estimates provided in the literature, which is consistent with the local context of a remote rural area of a developing country with few alternative online options available.

Our paper provides evidence for a new way to reduce this learning loss. Other strategies than online tutoring are currently being discussed to mitigate the learning loss of closing schools: SMS and 5-10-minute phone calls in Botswana (Angrist, Bergman and Matsheng, 2022), a similar program in Nepal (Radhakrishnan et al., 2021), 30-minute phone calls by teachers in Bangladesh (Beam, Mukherjee and Navarro-Sola, 2022), 30-minute phone tutoring sessions in Bangladesh (Hassan et al., 2022), teacher-student 15-minute mini-tutoring sessions in Kenya (Schueler and Rodriguez-Segura, 2021), and weekly phone tutorials from teachers in Sierra Leone (Crawfurd et al., 2022). The contribution of our paper is to study for the first time online video tutoring. We demonstrate that school closures led to significant learning loss (0.87 SD), 47% of which was compensated for by the tutoring program.

We also shed light on the heterogeneous effects of the tutoring program across time. While the program turned out to be a crucial part of students' education during lockdown, its impact on student grades in a normal time period was not statistically different from 0. This confirms the findings of a literature on tutoring that has found no effects when the schools are open (Nickow, Oreopoulos and Quan (2020) for non-professional volunteer tutors in after-school tutoring programs (the case in our paper), Romero, Chen and Magari (2021) for cross-age tutoring in Kenya, Ly, Maurin and Riegert (2020) in France, Kraft et al. (2022) for online tutoring in the US) but an

effect when the schools are closed (Carlana and La Ferrara, 2021). This result is perhaps not very surprising ex-post; the marginal returns to an additional hour of tutoring are likely to be high when students aren't receiving any other education, but may be low if they are attending school full-time.

A limitation of our study is external validity since our sample is small and the intervention is implemented in rural Kenya. Reassuringly, our results are firmly within those of the existing literature in various different contexts in Italy (Carlana and La Ferrara, 2021), Botswana (Angrist, Bergman and Matsheng, 2022), Nepal (Radhakrishnan et al., 2021), Bangladesh (Beam, Mukherjee and Navarro-Sola, 2022; Hassan et al., 2022), Kenya (Schueler and Rodriguez-Segura, 2021), and Sierra Leone (Crawfurd et al., 2022).

Overall, we conclude that online tutoring can recover almost half of the cognitive losses, but none of the losses in aspirations. School closures had profound effects that must be fully understood and carefully estimated before closing schools.

REFERENCES

Anderson, Michael L. 2008. "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association*, 103(484): 1481–1495.

Angrist, Noam, Peter Bergman, and Moitshepi Matsheng. 2022. "Experimental evidence on learning using low-tech when school is out." *Nature Human Behaviour*, 6: 941–950.

Azevedo, Joao Pedro, Amer Hasan, Diana Goldemberg, Syedah Aroob Iqbal, and Koen Geven. 2020. *Simulating the Potential Impacts of COVID-19 School Closures on Schooling and Learning Outcomes: A Set of Global Estimates.*

Banerjee, Abhijit, Esther Duflo, Amy Finkelstein, Lawrence Katz, Benjamin Olken, and Anja Sautmann. 2020. "In Praise of Moderation: Suggestions for the Scope and Use of Pre-Analysis Plans for RCTs in Economics." *NBER Working Papers 26993, National Bureau of Economic Research, Inc.*

Banerji, Rukmini, James Berry, and Marc Shotland. 2017. "The Impact of Maternal Literacy and Participation Programs: Evidence from a Randomized Evaluation in India." *American Economic Journal: Applied Economics*, 9(4): 303–37.

BBC. 2021. "Coronavirus: Kenya reopens schools after nine months."

BBC. 2022. "Uganda schools reopen after almost two years of Covid closure."

Beam, Emily, Priya Mukherjee, and Laia Navarro-Sola. 2022. "Lowering Barriers to Remote Education: Experimental Impacts on Parental Responses and Learning." *IZA DP No. 15596*.

Cameron, A Colin, Jonah B Gelbach, and Douglas L Miller. 2008. "Bootstrap-based improvements for inference with clustered errors." *The Review of Economics and Statistics*, 90(3): 414–427.

Carlana, Michela, and Eliana La Ferrara. 2021. "Apart but Connected: Online Tutoring and Student Outcomes during the COVID-19 Pandemic." *Working Paper*.

Chemin, Matthieu. 2018. "Informal Groups and Health Insurance Take-up Evidence from a Field Experiment." *World Development*, 101: 54–72.

Crawfurd, Lee, David Evans, Susannah Hares, and Justin Sandefur. 2022. "Live Tutoring Calls Did Not Improve Learning during the COVID-19 Pandemic in Sierra Leone." *CGD Working Paper 591*.

Education: From disruption to recovery. 2022.

Engzell, Per, Arun Frey, and Mark D. Verhagen. 2021. "Learning loss due to school closures during the COVID-19 pandemic." *Proceedings of the National Academy of Sciences*, 118(17): e2022376118.

Glewwe, P., and K. Muralidharan. 2016. "Chapter 10 - Improving Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications." In . Vol. 5 of *Handbook of the Economics of Education*, , ed. Eric A. Hanushek, Stephen Machin and Ludger Woessmann, 653–743. Elsevier.

Gore, Jennifer, Leanne Fray, Andrew Miller, Jess Harris, and Wendy Taggart. 2021. "The impact of COVID-19 on student learning in New South Wales primary schools: an empirical study." *The Australian Educational Researcher*, 48.

Hanushek, Eric A., and Ludger Woessmann. 2020. "The economic impacts of learning losses." *OECD*, , (225).

Hassan, Hashibul, Asad Islam, Abu Siddique, and Liang Choon Wang. 2022. "Telementoring and homeschooling during school closures: A randomized experiment in rural Bangladesh." *working paper*.

Hevia, Felipe J., Samana Vergara-Lope, Anabel Velásquez-Durán, and David Calderón. 2022. "Estimation of the fundamental learning loss and learning poverty related to COVID-19 pandemic in Mexico." *International Journal of Educational Development*, 88: 102515.

Hjort, Jonas, and Jonas Poulsen. 2019. "The Arrival of Fast Internet and Employment in Africa." *American Economic Review*, 109(3): 1032–79.

Kraft, Matthew A., John A. List, Jeffrey A. Livingston, and Sally Sadoff. 2022. "Online Tutoring by College Volunteers: Experimental Evidence from a Pilot Program." *AEA Papers and Proceedings*, 112: 614–618.

Kuhfeld, Megan, Beth Tarasawa, Angela Johnson, Erik Ruzek, and Karyn Lewis. 2020. "Learning during COVID-19: Initial findings on students' reading and math achievement and growth."

Ly, Son Thierry, Eric Maurin, and Arnaud Riegert. 2020. "A Pleasure That Hurts: The Ambiguous Effects of Elite Tutoring on Underprivileged High School Students." *Journal of Labor Economics*, 38(2): 501–533.

Mabeya, Mary Theodorah. 2020. "Distance Learning During COVID-19 Crisis: Primary and Secondary School Parents Experiences in Kenya." *East African Journal of Education Studies*, 2(1).

Maldonado, Joana, and Kristof De Witte. 2020. "The effect of school closures on standardised student test outcomes." *British Educational Research Journal*.

Malenya, Francis Likoye, and Asayo Ohba. 2023. "Equity issues in the provision of online learning duringthe Covid-19 pandemic in KenyaEquity issues during online learning in KenyaReceived 23 December 2022 Revised 5 January 2023 Accepted 5 January 2023." *Journal of International Cooperation in Education*.

McKenzie, David. 2012. "Beyond baseline and follow-up: The case for more T in experiments." *Journal of Development Economics*, 99: 201–221.

Moscoviz, Laura, and David Evans. 2022. "Learning Loss and Student Dropouts during the COVID-19 Pandemic: A Review of the Evidence Two Years after Schools Shut Down." *CGD Working Paper 609*.

Muris, Peter. 2001. "A Brief Questionnaire for Measuring Self-Efficacy in Youths." *Journal of Psychopathology and Behavioral Assessment*, 23: 145–149.

Nickow, Andre, Philip Oreopoulos, and Vincent Quan. 2020. "The Impressive Effects of Tutoring on PreK-12 Learning: ASystematic Review and Meta-Analysis of the Experimental Evidence." *Working Paper 27476 NBER.*

Ochieng, Vollan Okoth, and Moses Waithanji Ngware. 2022. "Adoption of Education Technologies for Learning During COVID-19 Pandemic: The Experiences of Marginalized and Vulnerable Learner Populations in Kenya." *International Journal of Educational Reform*, 1–24.

Oketch, Moses, and Maurice Mutisya. 2013. "Evolutional of educational outcomes in Kenya." *Paper commissioned for the EFA Global Monitoring Report 2013/4, Teaching and learning: Achieving quality for all.*

Olanrewaju, Gideon Seun, Seun Bunmi Adebayo, Abiodun Yetunde Omotosho, and Charles Falajiki Olajidea. 2021. "2021; 2: 100092.Published online 2021 Nov 18. doi: 10.1016/j.ijedro.2021.100092PMCID: PMC8600108PMID: 35059671Left behind? The effects of digital gaps on e-learning in rural secondary schools and remote communities across Nigeria during the COVID19 pandemic." *International Journal of Educational Research Open.*

Patrinos, Harry Anthony, Emiliana Vegas, and Rohan Carter-Rau. 2022. "An Analysis of COVID-19 Student Learning Loss." *World Bank Policy Research Working Paper 10033.*

Pell, Tony, and Tina Jarvis. 2001. "Developing attitude to science scales for use with children of ages from five to eleven years." *International Journal of Science Education*, 23(8): 847–862.

Radhakrishnan, Karthika, Shwetlena Sabarwal, Uttam Sharma, Claire Cullen, Colin Crossley, Thato Letsomo, and Noam Angrist. 2021. "Remote Learning: Evidence from Nepal during COVID-19." *World Bank Policy Brief.*

Romero, Mauricio, Lisa Chen, and Noriko Magari. 2021. "Cross-Age Tutoring: Experimental Evidence from Kenya." *Economic Development and Cultural Change.*

Rosenberg, Morris, Carmi Schooler, Carrie Schoenbach, and Florence Rosenberg. 1995. "Global Self-Esteem and Specific Self-Esteem: Different Concepts, Different Outcomes." *American Sociological Review*, 60(1): 141–156.

Schueler, Beth, and Daniel Rodriguez-Segura. 2021. "A Cautionary Tale of Tutoring Hard-to-Reach Students in Kenya." *EdWorkingPaper: 21-432.*

Schult, Johannes, Nicole Mahler, Benjamin Fauth, and Marlit A Lindner. 2021. "Did Students Learn Less During the COVID-19 Pandemic? Reading and Mathematics Competencies Before and After the First Pandemic Wave."

Singh, Abhijeet, Mauricio Romero, and Karthik Muralidharan. 2022. "COVID-19 Learning Loss and Recovery: Panel Data Evidence from India." *National Bureau of Economic Research Working Paper 30552.*

Webb, Matthew. 2014. "Reworking Wild Bootstrap Based Inference for Clustered Errors." *Economics Department, Queen's University.*

Woessmann, Ludger, Vera Freundl, Elisabeth Grewenig, Philipp Lergetporer, Katharina Werner, and Larissa Zierow. 2020. "Education in the corona crisis: How did the schoolchildren spend the time the schools were closed and which educational measures do the Germans advocate?" *ifo Institute.*

Young, Alwyn. 2018. "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results." *The Quarterly Journal of Economics*, 134(2): 557–598.

APPENDIX

Appendix A: English Proficiency Measure

To measure oral proficiency in English (which is not assessed in the exams), we use a test constructed to be mapped into international language standards. Native English speakers were hired and paid by an external organization (called the "McGill Arts Internship Office") to physically travel to the site of the research project in Kenya. These interns were not tutors themselves and were blind to the experiment in the sense that they were never shown the randomized list of who was in the treatment or control group.[18]

These interns were all native English speakers and were thus able to gauge oral proficiency in English. They ask seven questions to start and facilitate a conversation. The first questions are easy with concrete subjects and a familiar vocabulary (i.e., Do you prefer rice or ugali?), while the last questions are harder with more abstract subjects (i.e., Can you describe for me the meaning of the word kindness?).[19] Considering the range of questions, the test is designed to be informative over a wide range of student achievement.

These questions are different from those suggested for the ice-breaking activities of the tutors. The tutors were never informed about the content of this oral proficiency test such that it would not have been possible for them to teach to the test. In any case, tutors had no incentives to teach to the test, they were entirely volunteering their time with no rewards being given for certain results.

These questions were carefully chosen after extensive piloting to deal with issues of time and shyness. The reasoning behind them was that asking students more direct questions elicited more direct answers. In a previous version of the test, we showed cartoons and asked students to describe them, followed by a storytelling/listening activity. The open-endedness of the photo-based questions struck students silent – even those that spoke English well. After that, it was hard to refocus the conversation, and the interview became awkward. This obviously only made students clam up more. We discovered it was easier to ask a question, see what happens, and continue. The pictures

[18]The main occupation of these interns was to develop their own independent research project (different from this project), collect their own data, analyze it and produce a working paper for academic credits on their return to the university. To get experience collecting data, they collected these oral proficiency tests. These interns were not paid by the experimenter.

[19]The full list is: 1. What is your name? How old are you? Do you have any brothers or sisters? Can you tell me about them? (Finding out basic personal information, warm-up questions.) 2. Do you prefer rice or ugali? Why is that? (Warm-up, concrete subject, familiar vocabulary, likes/dislikes.) 3. Do you have a musician or television program? Can you tell me about it/them? Why do you like it/them? (Concrete subject, likes/dislikes, and opportunity to demonstrate range of vocabulary and fluency.) 4. Can you name a sport you would like to play one day? A food you would like to try? A place you would like to visit? (Concrete subject, less familiar vocabulary, uses future tense.) 5. Can you describe for me the meaning of the word kindness? (Abstract subjects.) 6. Can you think of an occasion where you were very happy? Can you tell me about it? (Abstract subjects, past tense.) 7. I want you to try to think of a question to ask me. It can be about anything! (Ability to ask questions.)

were overwhelming. It was also hard to find cartoons that both suited the context and had enough activity going on. The storytelling/listening activity made the test too long. The students have limited attention spans, and once they lost interest or sat in silence for too long, it was hard to get them back on track. The test used in this paper with a few direct questions deals with these issues of time and shyness. The beginning conversational questions get students comfortable and give them time to warm up. Having pictures to look at and things to listen to made it feel like more of a "test", whereas the few questions is more of a casual "chit chat." In this way, the native English speakers were able to elicit responses from students and gauge their level of oral proficiency.

The native English speakers then grade each student on four different dimensions: understanding a native speaker, conversation, vocabulary range, and spoken fluency. They use a "rubric", i.e., in education terminology, a scoring guide used to evaluate the quality of students' constructed responses, established by the "Common European Framework of Reference for Languages (CEFR)", put together by the Council of Europe as a way of standardizing the levels of language exams in different regions. The CEFR scoring rubrics are important since they are widely used internationally and all important exams are mapped to them.[20]

The rubrics for the Oral Proficiency Test are:

- Understanding a native speaker

6: Proficient Can understand any native speaker, even on abstract and complex topics of a specialist nature beyond his/her own field, given an opportunity to adjust to a non-standard accent or dialect.

5: Advanced Can understand in detail speech on abstract and complex topics of a specialist nature beyond his/her own field, though he/she may need to confirm occasional details, especially if the accent is unfamiliar.

4: Early Advanced Can understand in detail what is said to him/her in the standard spoken language even in a noisy environment.

3: Intermediate Can follow clearly articulated speech directed at him/her in everyday conversation, though will sometimes have to ask for repetition of particular words and phrases.

2: Early Intermediate Can understand enough to manage simple, routine exchanges without undue effort. Can generally understand clear, standard speech on familiar matters directed at him/her, provided he/she can ask for repetition or reformulation from time to time.

[20]See for more details: https://www.coe.int/en/web/common-european-framework-reference-languages/home

1: Beginning Can understand everyday expressions aimed at the satisfaction of simple needs of a concrete type, delivered directly to him/her in clear, slow and repeated speech by a sympathetic speaker. Can understand questions and instructions addressed carefully and slowly to him/her and follow short, simple directions.

- Conversation

6: Proficient Can converse comfortably and appropriately, unhampered by any linguistic limitations in conducting a full social and personal life.

5: Advanced Can use language flexibly and effectively for social purposes, including emotional, allusive and joking usage.

4: Early Advanced Can engage in extended conversation on most general topics in a clearly participatory fashion, even in a noisy environment. Can sustain relationships with native speakers without unintentionally amusing or irritating them or requiring them to behave other than they would with a native speaker. Can convey degrees of emotion and highlight the personal significance of events and experiences.

3: Intermediate Can enter unprepared into conversations on familiar topics. Can follow clearly articulated speech directed at him/her in everyday conversation, though will sometimes have to ask for repetition of particular words and phrases. Can maintain a conversation or discussion but may sometimes be difficult to follow when trying to say exactly what he/she would like to. Can express and respond to feelings such as surprise, happiness, sadness, interest and indifference.

2: Early Intermediate Can establish social contact: greetings and farewells; introductions; giving thanks. Can generally understand clear, standard speech on familiar matters directed at him/her, provided he/she can ask for repetition or reformulation from time to time. Can participate in short conversations in routine contexts on topics of interest. Can express how he/she feels in simple terms, and express thanks. Can handle very short social exchanges but is rarely able to understand enough to keep conversation going of his/her own accord, though he/she can be made to understand if the speaker will take the trouble. Can use simple everyday polite forms of greeting and address. Can make and respond to invitations, suggestions and apologies. Can say what he/she likes and dislikes.

1: Beginning Can make an introduction and use basic greeting and leave-taking expressions. Can ask how people are and react to news. Can understand everyday expressions aimed at the

satisfaction of simple needs of a concrete type, delivered directly to him/her in clear, slow and repeated speech by a sympathetic speaker.

- Vocabulary range

6: Proficient Has a good command of a very broad lexical repertoire including idiomatic expressions and colloquialisms; shows awareness of connotative levels of meaning.

5: Advanced Has a good command of a broad lexical repertoire allowing gaps to be readily overcome with circumlocutions; little obvious searching for expressions or avoidance strategies. Good command of idiomatic expressions and colloquialisms.

4: Early Advanced Has a good range of vocabulary for matters connected to his/her field and most general topics. Can vary formulation to avoid frequent repetition, but lexical gaps can still cause hesitation and circumlocution.

3: Intermediate Has a sufficient vocabulary to express him/herself with some circumlocutions on most topics pertinent to his/her everyday life such as family, hobbies and interests, work, travel, and current events.

2: Early Intermediate Has sufficient vocabulary to conduct routine, everyday transactions involving familiar situations and topics. Has a sufficient vocabulary for the expression of basic communicative needs. Has a sufficient vocabulary for coping with simple survival needs.

1: Beginning Has a basic vocabulary repertoire of isolated words and phrases related to particular concrete situations.

- Spoken fluency

6: Proficient Can express him/herself at length with a natural, effortless, unhesitating flow. Pauses only to reflect on precisely the right words to express his/her thoughts or to find an appropriate example or explanation.

5: Advanced Can express him/herself fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language.

4: Early Advanced Can communicate spontaneously, often showing remarkable fluency and ease of expression in even longer complex stretches of speech. Can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he/she searches for patterns and expressions, there are few noticeably long pauses. Can interact with a degree of fluency

and spontaneity that makes regular interaction with native speakers quite possible without imposing strain on either party.

3: Intermediate Can express him/herself with relative ease. Despite some problems with formulation resulting in pauses and 'cul-de-sacs', he/she is able to keep going effectively without help. Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production.

2: Early Intermediate Can make him/herself understood in short contributions, even though pauses, false starts and reformulation are very evident. Can construct phrases on familiar topics with sufficient ease to handle short exchanges, despite very noticeable hesitation and false starts.

1: Beginning Can manage very short, isolated, mainly pre-packaged utterances, with much pausing to search for expressions, to articulate less familiar words, and to repair communication.

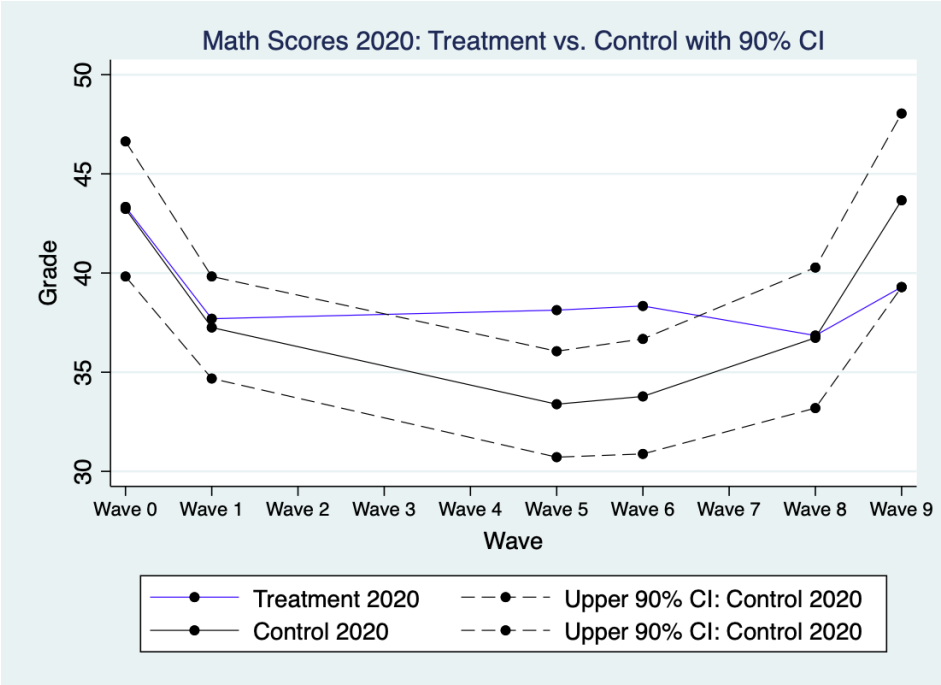
Appendix B: Confidence Intervals

in Figure 1, we display the confidence intervals for the control group in the year 2020. The average maths grade for the treatment group is above the upper bound of the 90% confidence interval in Waves 5 and 6, the waves directly following the reopening of schools.

The average maths grade for the treatment group remains within the confidence interval in Wave 9, indicating that the large spike observed for the control group in wave 9 is not significantly different from the treatment group.

We conclude that the only significant difference is in waves 5 and 6, not in other waves.

Figure B1. Math Grades: Treatment vs Control in times of COVID-19



Note: The figure shows trends across the 9 waves within the school year of 2020, split by treatment and control groups. The dashed lines indicate the 90% confidence interval for the control group. Schools were closed for waves 2 through 5 in 2020, but online tutoring continued.

Appendix C: Control Variables

Table C1 includes a student's baseline value for a specific index of survey questions. Specifically, we control for baseline English oral comprehension in Column (1), computer proficiency in Column (2), cross-culture communication in Column (3), motivation in Column (4), self-esteem in Column (5), aspiration in Column (6), liking school in Column (7), liking courses in Column (8), thoughts on Canada in Column (9), and thoughts on Kenya in Column (10) . We find similar results in all columns.

Table C1—Math grades: 2016-2018 vs 2020, Index Controls

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Dependent Variable: Math grade									
Math	-0.57	-0.65	-1.01	-0.43	-0.65	-0.59	-0.74	-0.68	-0.68	-0.66
	(1.25)	(1.20)	(1.21)	(1.18)	(1.20)	(1.23)	(1.24)	(1.23)	(1.20)	(1.20)
Math * School Closed	6.09**	5.50**	6.05**	5.53**	5.65**	5.60**	5.67**	5.67**	5.57**	5.65**
	(2.83)	(2.72)	(2.67)	(2.68)	(2.70)	(2.70)	(2.70)	(2.70)	(2.69)	(2.69)
School Closed	-8.31***	-12.88***	-10.54***	-13.43***	-12.10***	-12.47***	-12.05***	-12.07***	-11.67***	-12.19***
	(2.30)	(2.13)	(2.20)	(2.22)	(2.20)	(2.16)	(2.17)	(2.19)	(2.31)	(2.25)
Wave 5 * 2016-2018	-0.23	-0.22	-0.24	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22
	(1.12)	(1.13)	(1.13)	(1.12)	(1.13)	(1.13)	(1.13)	(1.13)	(1.13)	(1.12)
English	-1.43	-1.11	-1.12	-0.79	-1.03	-0.95	-1.06	-1.06	-1.06	-1.05
	(1.35)	(1.40)	(1.37)	(1.39)	(1.42)	(1.39)	(1.40)	(1.41)	(1.41)	(1.40)
Control	Oral Comp.	Computer Prof.	X-Culture Comm.	Motiv.	Self-Esteem	Aspiration	Like School	Like Courses	Canada	Kenya
Observations	2,170	2,170	2,170	2,170	2,170	2,170	2,170	2,170	2,170	2,170
R-squared	0.368	0.357	0.363	0.364	0.356	0.357	0.356	0.356	0.356	0.356
Mean Dep. Var	40.82	40.82	40.82	40.82	40.82	40.82	40.82	40.82	40.82	40.82
SD	13.77	13.77	13.77	13.77	13.77	13.77	13.77	13.77	13.77	13.77

Note: Standard errors clustered at the student level in parentheses.. *** p<0.01, ** p<0.05, * p<0.1. For each column, the dependent variable is a student's math grade in a given wave. "Math Treatment" is a dummy equal to 1 if a student is in the Math treatment group. "School Closed" is a dichotomous variable equal to 1 when schools are closed, i.e., wave 5 of 2020. "MathTreatment * School Closed" is the interaction between the two variables.. "Wave 5 * 2016-2018" is a dummy equal to 1 if the period is wave 5 in any of the years 2016 - 2018. "English" is a dummy variable equal to 1 student i was in the treatment group for the years 2016 to 2018. All regressions include a full set of interactions between the 9 waves and the 3 time periods (2016-2018 for the English treatment, 2019 for the Math treatment, and 2020 for the Math treatment combined with school closure). All regressions control for baseline math grade and a dummy variable "Baseline Missing" equal to 1 if the baseline math grade is missing. If baseline math grade is missing, it is replaced by the value 0 and the dummy variable "Baseline Missing" takes the value 1. Each column includes a student's baseline value for a specific index of survey questions. For columns 1-10, these indices include respectively: oral comprehension, computer proficiency, cross-culture communication, motivation, self-esteem, future aspirations, liking school in general, liking courses, thoughts about living in Canada, and thoughts about living in Kenya.

Table C2 includes a student's baseline value for the 51 components of the 10 sections of the survey.

Table C2—All Components

	Maths
Math	0.23
	(1.14)
Math * School Closed	4.85*
	(2.50)
School Closed	-28.26***
	(4.81)
English	-1.31
	(1.21)
Control Variable	All Components
Observations	2,113
R-squared	0.483
Mean Dep. Var.	40.82
SD	13.77

Note: Standard errors clustered at the student level in parentheses. *** p<0.01, ** p<0.05, * p<0.1. This column includes the baseline value of the 51 components of the 10 sections of the survey.

Appendix D: Disaggregating wave*year fixed effects

Our specification in Table 3 groups together the years 2016-2018 within the wave*year fixed effects. In Table D1 below, we use the same specification from Table 3- the only difference is that we relax the above assumption by disaggregating the wave*year fixed effects. This has essentially no impact on the results from Table 3. Column (1) of Table D1 shows that the variable $Math_{it}$ is still not statistically different from 0, indicating that the educational program did not have any significant effect on math grades. When the schools are closed, the coefficient of $Math * SchoolClosed_{it}$ is statistically significant at the 5% level and now has a coefficient of 5.87, indicating that being in the treatment group was associated with a math score 5.87 points higher compared to the control group.

Table D1—Math grades: Disaggregated Wave*Year Fixed Effects for the 2016-2018 period

	(1)	(2)	(3)	(4)
	Dependent Variable: Math grade			
Math	-0.84	-0.82	-0.59	-0.58
	(1.28)	(1.21)	(1.20)	(1.16)
Math * School Closed	5.87**	6.39**	6.54**	6.12**
	(2.76)	(2.72)	(2.79)	(2.77)
School Closed	-10.67***	-10.90***	-11.64***	-11.08**
	(2.11)	(2.23)	(2.44)	(4.35)
Wave*Year fixed effects	YES	YES	YES	YES
Controls:				
Baseline grade	NO	YES	YES	YES
Age, Gender, School Year	NO	NO	YES	YES
Baseline Survey	NO	NO	NO	YES
Observations	2,170	2,170	2,170	2,170
R-squared	0.537	0.560	0.564	0.572
Mean dep var	40.82	40.82	40.82	40.82
SD dep var	13.77	13.77	13.77	13.77

Note: Standard errors clustered at the student level in parentheses. *** p<0.01, ** p<0.05, * p<0.1. For each column, the dependent variable is a student's math grade in a given wave. 'Math' is a dummy equal to 1 if a student was in the Math treatment group for the years 2019-2020. 'School Closed' is a dichotomous variable equal to 1 when schools are closed, i.e., wave 5 of 2020. 'MathTreatment * School Closed' is the interaction between the two variables. All regressions include a full set of interactions between the 9 waves and the 5 years. All regressions control for baseline math grade and a dummy variable 'Baseline Missing' that is equal to 1 if the baseline math grade is missing. If baseline math grade is missing, it is replaced by the value 0 and the dummy variable 'Baseline Missing' takes the value 1. Column 1 shows the estimation of Equation 1 without any additional controls. Column 2 augments this specification by including the baseline grade on all other topics than Maths, and a dummy variable 'Baseline Missing Total grade' equal to 1 if the baseline total grade is missing. Column 3 adds to column 2 by controlling for a student's age, gender, and the year of schooling they are currently completing. Finally, column 4 includes in the list of controls the baseline averages of various indices from the survey data. These indices include: oral comprehension, computer proficiency, cross-culture communication, motivation, self-esteem, future aspirations, liking school, liking classes, thoughts about Canada, and thoughts about Kenya.

Appendix E : Estimation with 2019-2020

In Table E1 below, we restrict the sample to the years 2019 and 2020 alone when the tutoring was in Maths. The results are very similar to the main results of the paper.

Table E1—Math Grades: 2019 and 2020

	(1)	(2)	(3)	(4)
	Dependent Variable: Math Grade			
Math	-0.68	-0.53	-0.09	-0.44
	(1.20)	(1.09)	(1.13)	(1.04)
Math * School Closed	5.66**	6.10**	6.30**	6.31**
	(2.71)	(2.64)	(2.74)	(2.70)
School Closed	-7.31***	-7.15***	-6.51***	-8.14**
	(1.87)	(1.87)	(1.94)	(3.50)
Wave*Year fixed effects	YES	YES	YES	YES
Controls:				
Baseline Grade	NO	YES	YES	YES
Age, Gender, School Year	NO	NO	YES	YES
Baseline Survey	NO	NO	NO	YES
Observations	867	867	867	867
R-squared	0.439	0.481	0.488	0.503
Mean Dep. Var	37.51	37.51	37.51	37.51
SD	12.68	12.68	12.68	12.68

Note: Standard errors clustered at the student level in parentheses. *** p<0.01, ** p<0.05, * p<0.1. For each column, the dependent variable is a student's math grade in a given wave. 'Math' is a dummy equal to 1 if a student was in the Math treatment group for the years 2019-2020. 'School Closed' is a dichotomous variable equal to 1 when schools are closed, i.e., wave 5 of 2020. 'MathTreatment * School Closed' is the interaction between the two variables. All regressions include a full set of interactions between the 9 waves and the 2 time periods (2019 and 2020). All regressions control for baseline math grade and a dummy variable 'Baseline Missing' that is equal to 1 if the baseline math grade is missing. If baseline math grade is missing, it is replaced by the value 0 and the dummy variable 'Baseline Missing' takes the value 1. Column 1 shows the estimation of Equation 1 without any additional controls. Column 2 augments this specification by including the baseline grade on all other topics than Maths, and a dummy variable 'Baseline Missing Total Grade' equal to 1 if the baseline total grade is missing. Column 3 adds to column 2 by controlling for a student's age, gender, and the year of schooling they are currently completing. Finally, column 4 includes in the list of controls the baseline averages of various indices from the survey data. These indices include: oral comprehension, computer proficiency, cross-culture communication, motivation, self-esteem, future aspirations, liking school, liking classes, thoughts about Canada, and thoughts about Kenya.

Appendix F: Attrition Test

We create a dummy variable that represents a student's attrition status for a given wave-year. If the student is missing the math grade, the variable is set to 1. We first show that the group of students receiving the Math intervention is not associated with a statistically different attrition status in Column (1). We then add in the "School Closed" dummy and its interaction with the Math intervention in Column (2) , as well as baseline total grade, age, gender, school year, and baseline survey responses. Reassuringly, neither the School Closed period nor the interaction term are associated with higher or lower attrition.

Table F1—Attrition Test

	(1)	(2)
	Dependent Variable: Attrition	
Math	-0.00	-0.01
	(0.01)	(0.01)
Math * School Closed		-0.06
		(0.04)
School Closed		0.03
		(0.04)
Wave 5 * 2016-2018	-0.01	-0.01
	(0.02)	(0.02)
English	-0.03**	-0.02**
	(0.01)	(0.01)
Wave*Year fixed effects	YES	YES
Controls:		
Baseline grade	NO	YES
Age, Gender, School Year	NO	YES
Baseline Survey	NO	YES
Observations	2,212	2,212
R-squared	0.040	0.105
Mean dep var	0.019	0.019
SD dep var	0.138	0.138

Note: Standard errors clustered at the student level in parentheses. *** p<0.01, ** p<0.05, * p<0.1. For each column, the dependent variable is a student's attrition status in a given wave. 'Math' is a dummy equal to 1 if a student was in the Math treatment group for the years 2019-2020. 'School Closed' is a dichotomous variable equal to 1 when schools are closed, i.e., wave 5 of 2020. 'MathTreatment * School Closed' is the interaction between the two variables. 'Wave 5 * 2016-2018' is a dummy equal to 1 if the period is wave 5 in any of the years 2016 - 2018. 'English' is a dummy variable equal to 1 if a student was in the treatment group for the years 2016 - 2018. All regressions include a full set of interactions between the 9 waves and the 3 time periods (2016-2018 for the English treatment, 2019 for the Math treatment, and 2020 for the Math treatment combined with school closure).

Appendix G: Other Outcomes

Computer Proficiency

In the Computer Proficiency index, we ask students a series of questions about their comfort with using computers and technology. These include: "How comfortable do you feel using a computer, including the internet?"; "How comfortable do you feel using the internet on a computer?"; "How comfortable do you feel using the internet on a cell phone?"; "How comfortable do you feel sending an email?"; "How comfortable do you feel talking on Skype?" All questions range from 1 to 5, with 5 being the highest comfort level.

Table G1—Computer Proficiency

	(1)	(2)	(3)	(4)	(5)	(6)
	Index	Using computer	Using Internet on computer	Using internet on phone	Using email	Using video call
Math	0.80***	0.84***	0.85***	0.42***	0.43***	1.75***
	(0.11)	(0.15)	(0.15)	(0.10)	(0.11)	(0.15)
Math * School Closed	0.31	0.31	0.30	0.23	0.54**	-0.27
	(0.20)	(0.27)	(0.26)	(0.20)	(0.23)	(0.21)
School Closed	0.87***	0.66**	1.08***	1.46***	0.84***	1.03***
	(0.25)	(0.27)	(0.37)	(0.28)	(0.29)	(0.34)
English	0.87***	0.96***	0.43***	0.22**	0.02	2.69***
	(0.06)	(0.11)	(0.09)	(0.10)	(0.04)	(0.12)
Observations	819	811	808	811	803	604
R-squared	0.545	0.285	0.443	0.586	0.440	0.648
Mean Dep. Var	2.228	2.476	1.996	2.308	1.445	3.280
SD	0.976	1.199	1.206	1.313	0.843	1.392

Note: Standard errors clustered at the student level in parentheses.. *** p<0.01, ** p<0.05, * p<0.1. Each component goes from 1 to 5, with 5 being the most comfort. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Column 1 shows the estimation of equation 1 with the unweighted average of the computer proficiency index as the dependent variable. Columns 2 through 5 show the results of the same regression specification but with each individual component of the computer proficiency index as the dependent variable. The components of this index all range from 1 to 5, with 1 indicating the least amount of comfort and 5 indicating the highest level of comfort. They include: using a computer, using internet on a computer, using internet on a phone, using email, and using video call. All regressions control for a student's age, gender, and current year of schooling, as well as the baseline survey response of the dependent variable and a dummy variable "Baseline Missing" equal to 1 if the baseline survey response is missing. If baseline survey response is missing, it is replaced by the value 0 and the dummy variable "Baseline Missing" takes the value 1.

Table G1 shows the estimation of equation 1 with the unweighted average of the computer proficiency index as the dependent variable, followed by each individual question. Both interventions (Math and English) significantly increase the score, by a large amount (0.8 and 0.87 out of 5). Notice that the coefficient of SchoolClosed is positive. One potential explanation for this seem-

ingly counterintuitive result is that the positive coefficient is a reflection of the general trend that students become more proficient with technology over time. Indeed, other (omitted) wave * year fixed effects also show positive coefficients. Since the omitted wave is "Wave 1 * 2016-2018" (i.e., the very first wave in the sample), all other wave * year fixed effects capture student-invariant time trends later in time.

Liking School

Table G2 shows the components of the Liking School index. We use modified questions from Pell and Jarvis (2001) and asked students how they felt about doing or learning certain subjects in school. In each column, the question asked to students is: "How do you feel about learning this subject/doing this activity related to school?", where student answers vary from 1 (don't like at all) to 5 (really like). They include: English composition, learning English, mathematics, christian religion, social studies, science, insha, and swahili.

School closures improve the scores for almost each field. In other words, students declare that they like school when schools are closed.

The interaction of the math tutoring program and school closures has no impact on liking school: the tutoring program does not have a differential effect on the Liking School index over and above the school closures.

Table G3 shows the results of these regressions for the Liking School index. We use an unweighted average index of all subjects in Column 1: school closures improve the liking school index.

In Column 2, we ask: "How do you feel about working by yourself at school?". When the schools are closed, students report liking less working alone. Column 3 shows on the other hand that students prefer working with others. Column 4 shows that student feel better about coming to school (the exact question is: "How do you feel about coming to school?", once again when the schools are closed.

Intuitively, the two results on loss of aspirations and liking school go hand-in-hand: students like school, but schools are closed - this hurts students' aspirations because they know it will be harder to go to university and get high-skilled jobs.

Table G2—Liking School

	(1) English Comp	(2) Learning English	(3) Math	(4) Christian Religion	(5) Social Studies	(6) Science	(7) Insha	(8) Swahili
Math	-0.05	-0.01	-0.03	-0.12*	-0.03	0.01	0.05	0.12
	(0.09)	(0.08)	(0.10)	(0.06)	(0.11)	(0.11)	(0.09)	(0.08)
Math * School Closed	0.02	-0.07	0.12	0.40***	-0.09	-0.20	-0.28	-0.09
	(0.23)	(0.21)	(0.28)	(0.15)	(0.23)	(0.24)	(0.21)	(0.19)
School Closed	0.32*	0.27	0.11	0.48***	0.18	0.40***	0.72***	0.58***
	(0.17)	(0.17)	(0.21)	(0.12)	(0.16)	(0.15)	(0.14)	(0.14)
English	-0.01	0.18**	-0.04	-0.07	0.02	-0.03	-0.09	0.06
	(0.08)	(0.09)	(0.09)	(0.08)	(0.09)	(0.07)	(0.08)	(0.07)
Observations	820	821	822	820	822	822	822	821
R-squared	0.200	0.140	0.094	0.193	0.116	0.129	0.164	0.192
Mean Dep. Var	3.830	4.041	3.912	4.045	3.691	4.030	3.811	4.004
SD	0.755	0.762	0.829	0.682	0.840	0.737	0.782	0.748

Note: Standard errors clustered at the student level in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Column 1 shows the estimation of equation 1 with the unweighted average of the self-esteem index as the dependent variable. Columns 2-9 go from 1-5. In each column, the question asked to students is: "How do you feel about learning this subject/doing this activity related to school?", where student answers vary from 1 (don't like at all) to 5 (really like). They include: English composition, learning English, mathematics, christian religion, social studies, science, insha, and swahili. All regressions control for a student's age, gender, and current year of schooling, as well as the baseline survey response of the dependent variable and a dummy variable "Baseline Missing" equal to 1 if the baseline survey response is missing. If baseline survey response is missing, it is replaced by the value 0 and the dummy variable "Baseline Missing" takes the value 1.

Table G3—Liking School

	(1)	(2)	(3)	(4)
	Index	Working alone	Working with others	Coming to school
Math	-0.05	-0.14*	-0.29***	-0.12**
	(0.04)	(0.09)	(0.10)	(0.05)
Math * School Closed	0.02	0.26**	0.39**	0.00
	(0.11)	(0.11)	(0.17)	(0.11)
School Closed	0.23***	-1.68***	0.48***	0.76***
	(0.07)	(0.13)	(0.13)	(0.07)
English	0.01	0.04	0.11	-0.01
	(0.03)	(0.09)	(0.07)	(0.04)
Observations	821	821	822	820
R-squared	0.233	0.515	0.200	0.315
Mean Dep. Var	3.897	3.028	4.133	4.343
SD	0.362	1.136	0.854	0.538

Note: Standard errors clustered at the student level in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Column 1 shows the estimation of equation 1 with the unweighted average of the liking school index as the dependent variable. For each subject, the question asked to students is: "How do you feel about learning this subject/doing this activity related to school?", where student answers vary from 1 (don't like at all) to 5 (really like). They include: English composition, learning English, mathematics, christian religion, social studies, science, insha, and swahili. In Column 2, the dependent variable is the answer to the question: "How do you feel about working by yourself at school?". In Column 3, the question is "How do you feel about working with others at school?", and in column 4 "How do you feel about coming to school?". Student answers vary from 1 (don't like at all) to 5 (really like). All regressions control for a student's age, gender, and current year of schooling, as well as the baseline survey response of the dependent variable and a dummy variable "Baseline Missing" equal to 1 if the baseline survey response is missing. If baseline survey response is missing, it is replaced by the value 0 and the dummy variable "Baseline Missing" takes the value 1.

Motivation

We use questions from Muris (2001) for the section on academic motivations, where each component in the Motivation index asks the student how well he or she can do on a certain task related to motivation (i.e. column 2 asks "How well can you get help when stuck on homework?"). The components all range from 1 to 5, with 1 indicating a very low ability to complete the task and 5 indicating a very strong ability. They include: getting help when stuck on homework, studying when there are other interesting things, doing revision before an exam, succeeding in finishing all your homework everyday, paying attention during every class, succeeding in passing courses, parents being satisfied with school performance, and easily passing a test.

Table G4—Motivation

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Index	Get help	Study	Revision	Finish homework	Pay attention	Passing courses	School performance	Pass test
Math	0.05	-0.01	0.06	0.05	0.09	0.03	0.03	0.09	-0.01
	(0.06)	(0.10)	(0.12)	(0.12)	(0.12)	(0.09)	(0.09)	(0.11)	(0.10)
Math * School Closed	0.11	0.11	-0.15	0.24	0.02	0.09	0.03	0.25	0.25
	(0.11)	(0.20)	(0.24)	(0.20)	(0.16)	(0.15)	(0.17)	(0.21)	(0.18)
School Closed	0.29***	-0.98***	-0.37**	-0.24	0.43***	0.16	0.96***	0.78***	1.82***
	(0.11)	(0.20)	(0.17)	(0.19)	(0.15)	(0.13)	(0.15)	(0.17)	(0.18)
English	0.01	0.04	-0.08	0.05	0.05	-0.06	-0.07	0.04	0.04
	(0.05)	(0.14)	(0.10)	(0.09)	(0.09)	(0.08)	(0.07)	(0.09)	(0.08)
Observations	821	819	819	820	822	822	822	821	821
R-squared	0.345	0.316	0.135	0.096	0.119	0.107	0.483	0.420	0.652
Mean Dep. Var.	3.258	2.938	2.573	3.840	3.791	3.940	3.294	2.903	2.781
SD	0.528	1.155	0.960	0.795	0.779	0.712	0.882	1.084	1.254

Note: Standard errors clustered at the student level in parentheses.. *** p<0.01, ** p<0.05, * p<0.1. Each component goes from 1 to 5, with 5 being the most, and each question asks "How well can you...." Column 1 shows the estimation of equation 1 with the unweighted average of the motivation index as the dependent variable. Columns 2 and 3 show the results of the same regression specification but with each individual component of the motivation index as the dependent variable. Each component in this index asks the student how well he or she can do on a certain task related to motivation (i.e. column 2 asks "How well can you get help when stuck on homework?"). The components all range from 1 to 5, with 1 indicating a very low ability to complete the task and 5 indicating a very strong ability, and include: getting help when stuck on homework, studying when there are other interesting things, doing revision before an exam, succeeding in finishing all your homework every day, paying attention during every class, succeeding in passing courses, parents being satisfied with your school performance, and easily passing a test. All regressions control for a student's age, gender, and current year of schooling, as well as the baseline survey response of the dependent variable and a dummy variable "Baseline Missing" equal to 1 if the baseline survey response is missing. If baseline survey response is missing, it is replaced by the value 0 and the dummy variable "Baseline Missing" takes the value 1.

Table G4 displays the results for the academic motivations module in the survey where students were asked this series of questions related to their school habits. Despite documenting an overall

positive effect of school closures on academic motivation, we find varying results across the components of the index: some components are positively affected (columns 5, 7, 8, 9) while others are negatively affected (columns 2 and 3). It is thus difficult to conclude that motivation is affected in a single direction.

Additionally, the interpretation of some of the questions is challenging during the school closure period. For example, students report struggling more during school closures with getting help when stuck on homework while also reporting that they succeed more in passing their courses during the school closures. Yet, there were no homework assignments or school tests during the period in which schools were closed.

Self-Esteem

Next, we use questions from Rosenberg et al. (1995) related to student self-esteem. The statements are: "I am satisfied with myself", "I think I am no good at all", "I feel that I have a number of good qualities", "I am able to do things as well as most others", "I feel I do not have much to be proud of", "I certainly feel useless at times", "I feel that I am a person of worth", "I wish I could have more respect for myself", "I sometimes feel that I'm a failure", and "I take a positive attitude toward myself". Answers range from 1 to 4, with 4 being strongly agree and 1 strongly disagree. Because columns 3, 6, 7, 9, and 10 ask questions where the ideal response is the lowest possible value, we reverse the response values (i.e. a response of 4 out of 4 for the question in column 6 now indicates that a student thinks he or she has much to be proud of). We calculate an unweighted average of these questions to build an index.

Column 1 of Table G5 seems to show that school closure is associated with overall higher student self-esteem, yet some individual components of the index show a positive sign (columns 2, 3, 4, 5, 7) and others show a negative sign (columns 6, 9) while still others show no effect (columns 8, 10, 11). It is thus difficult to conclude that self-esteem is unambiguously affected in a single direction.

Table G5—Self-Esteem

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Index	Satisfied	no good (inverse)	qualities	able	not proud (inverse)	useless (inverse)	worth	more respect (inverse)	failure (inverse)	positive
Math	-0.03	-0.04	-0.04	-0.02	0.01	-0.08	-0.00	-0.03	-0.03	-0.08	-0.07*
	(0.02)	(0.06)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.04)	(0.03)	(0.05)	(0.04)
Math * School Closed	-0.00	0.10	0.03	0.02	-0.05	0.03	-0.01	-0.12	0.07	0.01	-0.00
	(0.05)	(0.13)	(0.13)	(0.12)	(0.12)	(0.12)	(0.13)	(0.08)	(0.10)	(0.08)	(0.09)
School Closed	0.12***	0.83***	0.45***	0.18*	0.31***	-0.79***	0.23**	0.02	-0.20**	0.02	0.01
	(0.04)	(0.10)	(0.10)	(0.10)	(0.10)	(0.11)	(0.10)	(0.09)	(0.10)	(0.09)	(0.09)
English	0.02	0.02	0.02	-0.00	0.04	-0.04	0.12*	-0.02	-0.07	0.05	0.02
	(0.03)	(0.06)	(0.07)	(0.06)	(0.05)	(0.06)	(0.06)	(0.05)	(0.04)	(0.07)	(0.05)
Observations	821	820	821	815	821	804	821	821	814	816	820
R-squared	0.135	0.286	0.114	0.105	0.116	0.388	0.068	0.084	0.195	0.049	0.122
Mean Dep. Var	2.964	3.160	3.001	3.226	3.185	2.515	3.076	3.201	1.905	3.094	3.262
SD	0.263	0.633	0.608	0.463	0.522	0.674	0.524	0.439	0.406	0.526	0.467

Note: Standard errors clustered at the student level in parentheses.. *** p<0.01, ** p<0.05, * p<0.1. Column 1 shows the estimation of equation 1 with the unweighted average of the self-esteem index as the dependent variable. Columns 2 and 3 show the results of the same regression specification but with each individual component of the self-esteem index as the dependent variable. The components all range from 1 to 4, with 4 being strongly agree and 1 strongly disagree. They include: "I am satisfied with oneself", "I think I am no good at all", "I feel that I have a number of good qualities", "I am able to do things as well as most others", "I feel I do not have much to be proud of", "I certainly feel useless at times", "I feel that I am a person of worth", "I wish I could have more respect for myself", "I sometimes feel that I'm a failure", and "I take a positive attitude toward myself". Because columns 3, 6, 7, 9, and 10 ask questions where the ideal response is the lowest possible value, we reverse the response values (i.e. a response of 5 out of 5 for the question in column 6 now indicates that a student thinks he or she has much to be proud of). All regressions control for a student's age, gender, and current year of schooling, as well as the baseline survey response of the dependent variable and a dummy variable "Baseline Missing" equal to 1 if the baseline survey response is missing. If baseline survey response is missing, it is replaced by the value 0 and the dummy variable "Baseline Missing" takes the value 1.

Perceptions on Canada and Kenya

Finally, we present the results for the modules related to perceptions on Canada. The statement is: "Canada is a great place to live" and "Canada is a great place to be". The responses range from 1 to 4 for the first question, with 1 being strongly disagree and 4 being strongly agree, and 1 to 5 for the second question. In order to avoid the second question weighing more than the first due to a larger scale of possible responses, we rescale each question to range from 0 to 1. We build an index which is the unweighted average of these two questions.

The results are unclear in Table G6. The same analysis for Kenya in Table G7 tends to show that students disagreed more on average with the idea that Kenya is a great place to live during the period in which schools were closed.

Table G6—Canada

	(1)	(2)	(3)
	Index	Canada great place to live	Canada very good place to be
Math	-0.06***	-0.06	-0.26**
	(0.02)	(0.12)	(0.12)
Math * School Closed	0.02	0.60*	-0.46*
	(0.06)	(0.31)	(0.26)
School Closed	0.01	-0.85***	0.17
	(0.04)	(0.21)	(0.15)
English	0.00	-0.09	0.11
	(0.01)	(0.07)	(0.08)
Observations	821	736	820
R-squared	0.097	0.170	0.106
Mean Dep. Var	0.921	3.268	4.522
SD	0.148	0.747	0.824

Note: Standard errors clustered at the student level in parentheses.. *** p<0.01, ** p<0.05, * p<0.1. The table shows the estimation of equation 1. In column 1, the dependent variable is the unweighted average of the index on Canada. Columns 2 and 3 represent the components of this index. In column 2, the dependent variable represents the responses of students to the statement "Canada is a great place to live." The responses range from 1 to 4, with 1 being strongly disagree and 4 being strongly agree. Likewise, the dependent variable in column 3 is the response of students to the statement "Canada is a great place to be." The responses range from 1 to 5, with 1 being "strongly disagree" and 5 being "strongly agree".

Table G7—Kenya

	(1)	(2)	(3)
	Index	Kenya great place to live	Kenya very good place to be
Math	-0.02	-0.15	-0.03
	(0.01)	(0.10)	(0.13)
Math * School Closed	0.03	-0.27	0.43
	(0.03)	(0.20)	(0.27)
School Closed	-0.08***	0.09	-0.63***
	(0.02)	(0.12)	(0.22)
English	0.02	0.11*	0.01
	(0.01)	(0.06)	(0.11)
Observations	821	822	817
R-squared	0.174	0.139	0.143
Mean Dep. Var	0.857	3.575	3.998
SD	0.122	0.644	0.994

Note: Standard errors clustered at the student level in parentheses. *** p<0.01, ** p<0.05, * p<0.1. The table shows the estimation of equation 1. In column 1, the dependent variable is the unweighted average of the index on Kenya. Columns 2 and 3 represent the components of this index. In column 2, the dependent variable represents the responses of students to the statement "Kenya is a great place to live." The responses range from 1 to 4, with 1 being strongly disagree and 4 being strongly agree. Likewise, the dependent variable in column 3 is the response of students to the statement "Kenya is a great place to be." The responses range from 1 to 5, with 1 being "strongly disagree" and 5 being "strongly agree".

Appendix H: Validity and Reliability of Psychometric Tests

Below we present the tests of internal reliability, test-retest reliability, convergent validity, divergent validity and predictive validity.

Starting with internal reliability, Table H1 displays in column (1) the Alpha Cronbach test of each psychometric scales. The Alpha of the Oral Comprehension, Cross-Cultural Communication, Computer Proficiency, Liking School, Liking Courses, Academic Motivation, and Self-Esteem are all above 0.6, as per the NIH guidelines.[21] This indicates that the items composing a scale are well correlated with each other; i.e., when participants give a high response for one of the items, they are also likely to provide high responses for the other items. The Aspirations scale has a slightly lower alpha (0.51), not far from the guideline of 0.6. The alpha for the scales Thoughts on Canada and Thoughts on Kenya are 0.41 and 0.40. These scales are less central to the analysis, with only a remote link between tutoring and thoughts on Canada and Kenya; indeed we did not detect any meaningful treatment effect for these scales.

Table H1—Internal Reliability

	(1)	(2)
	Alpha Cronbach	ICC
Oral Comprehension	0.96	0.48**
Cross-Cultural Communication	0.68	0.28**
Computer Proficiency	0.87	0.34**
Aspirations	0.51	0.33**
Liking School	0.70	0.17**
Liking Courses	0.70	0.21**
Academic Motivation	0.64	0.45**
Self-Esteem	0.71	0.28**
Thoughts on Canada	0.41	0.02**
Thoughts on Kenya	0.40	0.28**

For test-retest reliability, we calculate the correlation between repeated waves for the same students. The intraclass correlation is displayed in Column (2). It is above 0.3 for most scales, and always statistically significant at 5 percent.

Convergent validity states that scales measuring the same concepts should positively correlate with each other. In Table H2, we measure the correlation between Oral Comprehension, Cross-Cultural Communication, Computer Proficiency, Aspirations, Liking School, Liking Courses, Aca-

[21] https://www.ncbi.nlm.nih.gov/books/NBK581902/

demic Motivation and Self-Esteem. The basic intuition is that these scales should be positively correlated, e.g., students motivated in class should also like courses. Indeed we find a positive correlation between all these scales, as displayed in the table. The only exception is the correlation between aspirations and proficiency with computer, which is negative. One may argue that these two concepts are not obviously connected, therefore a low correlation may be expected. In other words, one can be good at computers and have low aspirations, or vice versa.

Table H2—Convergent Validity

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Oral Comp.	Cross-Cultural Communication	Computer Proficiency	Aspirations	Liking School	Liking Courses	Academic Motivation	Self Esteem
Oral Comprehension	1							
Cross-Cultural Communication	0.39**	1						
Computer Proficiency	0.31**	0.44**	1					
Aspirations	0.09**	0.06**	-0.07**	1				
Liking School	0.14**	0.25**	0.20**	0.11**	1			
Liking Courses	0.19**	0.27**	0.29**	0.10**	0.95**	1		
Academic Motivation	0.20**	0.21**	0.30**	0.20**	0.45**	0.50**	1	
Self-Esteem	0.14**	0.16**	0.03	0.11**	0.25**	0.22**	0.38**	1

Divergent validity states that there should be no correlation between measures that should not have a relationship. To test for divergent validity, we focus on the scales Thoughts on Canada and Thoughts on Kenya. These scales ask about perceptions on the countries of Canada and Kenya (e.g., Canada is a great place to live; Kenya is a great place to live). These scales should not be connected with academic motivation or liking school; and indeed they are not, as Table H3 shows.

Finally, predictive validity evaluates how well a scale predicts an outcome. We use grades in school as an outcome in Table H4 below, and find that indeed the psychometric scales of Oral communication, Cross-cultural communication, Liking School, Liking courses, Academic Motivation, and Self Esteem are positively correlated with grades in school. The scale Aspirations is positively correlated with grades, very close to being significant. The scale Proficiency with computer is not correlated with grades, which may be expected since these are different skills. One can be proficient with using a computer, sending emails, but this does not necessarily correlate with grades.

Overall, we find that the psychometric scales used in this paper display internal reliability, test-retest reliability, convergent validity, divergent validity and predictive validity.

Table H3— Divergent Validity

	(1)	(2)
	Thoughts on Canada	Thoughts on Kenya
Oral communication	0.0666*	-0.1520*
Cross-cultural communication	-0.0084	-0.0623*
Proficiency with computer	0.0011	-0.1810*
Aspirations	0.1235*	0.1108*
Liking School	0.0136	0.0606*
Liking courses	0.0385	-0.0176
Academic Motivation	0.0431	-0.0599
Self Esteem	-0.0078	0.0764*

Table H4—Predictive Validity

	(1)
	Grade Total
Oral communication	0.2189*
Cross-cultural communication	0.0801*
Proficiency with computer	-0.0059
Aspirations	0.0635
Liking School	0.0906*
Liking courses	0.0982*
Academic Motivation	0.1275*
Self Esteem	0.0703*

Appendix I: Estimating the Learning Loss

We found that the online tutoring increases grades in Math when the schools are closed, but not when they are open. We use this fact to propose a methodology to estimate the learning loss, other than with a difference-in-differences.

We can summarize the three variables $Math$, $SchoolClosed$, and $Math * SchoolClosed$ into one single variable: the number of hours spent studying mathematics. When $Math = 0$ and $SchoolClosed = 0$, the student is in the control group and the schools are open. In that case, the students receives 3 hours of Math per week (which is the regular teaching load in Math for grade 6 students in Kenya). When $Math = 1$ and $SchoolClosed = 0$, the student is treated and receives an additional hour of Math per week.[22]

When $Math = 0$ and $SchoolClosed = 1$, the student receives no intervention and the schools are closed, such that the student receives no education in Math. Finally, when $Math = 1$ and $SchoolClosed = 1$, the student receives an hour of Math per week (intervention and schools closed).

Therefore, we construct a variable $MathHours_{itk}$ as the total number of hours per week student i spends studying mathematics from schooling and tutoring. According to the logic above, it is equal to 3 for the control group when the schools are open, 4 for the treatment group when the schools are open, 0 for the control group when the schools are closed, and 1 for the treatment group when the schools are closed.

Figure I1 below already lets on the idea of decreasing returns to math hours. We find a treatment effect when the schools are closed (for the first hour of Math taught) and no effect when the schools are open (moving from 3 to 4 hours of math, in fact a slightly negative effect but not significant). We superimpose a quadratic fitted line that clearly shows decreasing returns.

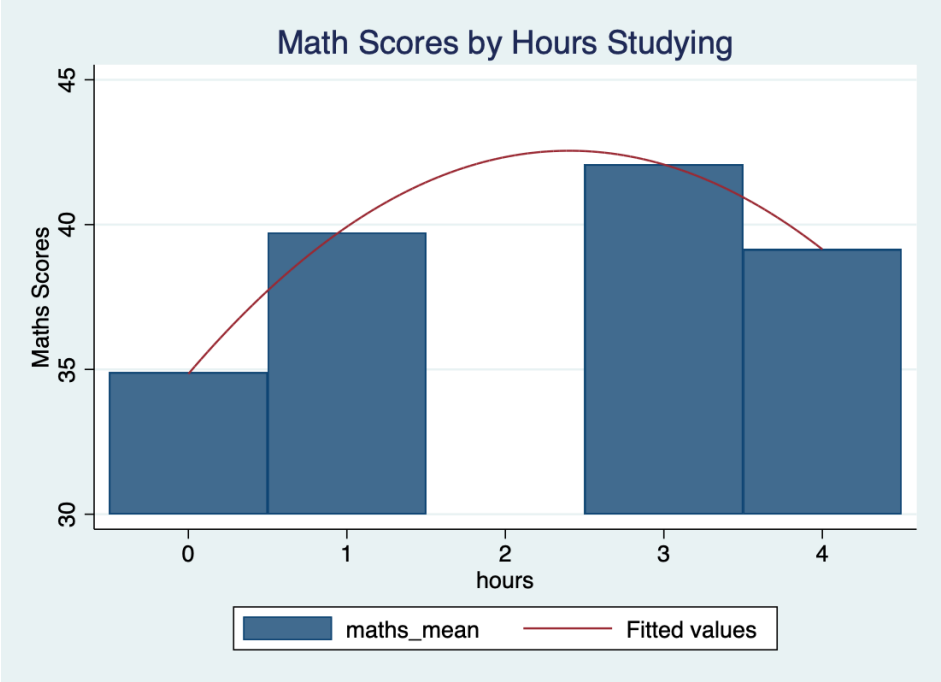
To capture these decreasing returns, we regress the math grade on Math hours and its squared term in the following specification:

$$y_{itk} = \beta_1 MathHours_{itk} + \beta_2 MathHours^2_{itk} + \beta_3 English_{it} + \beta_4 BaselineMathGrade_{it0} + \beta_5 BaselineMissing_{itk} + \beta_6 X_{it} + \beta_7 \delta_{tk} + \varepsilon_{itk},$$

where y_{itk} represents student i's math grade in year t, wave k. $MathHours_{itk}$ is the total number of hours per week student i spends studying mathematics from schooling and tutoring.

[22] One hour of tutoring may not be exactly comparable to one hour of teaching in class. One hour of tutoring may be more than one hour in class since the tutor is teaching one on one as opposed to the teacher teaching to an entire class. One hour of tutoring may be less if there are small interruptions or departures from the tutoring, such as when the tutor tries to get to know the tutee better through regular conversation, or occasional issues regarding the video quality. In any case, the relevant comparison in our analysis is the effect on Math grades with one hour of tutoring after 3 hours of teaching (when the schools are open) and after zero hours of teaching (when the schools are closed). That extra hour of tutoring is comparable. We repeated the analysis assuming that an extra hour of tutoring was equivalent to more or less of an hour in class, and find very similar results.

Figure I1. Math Hours



The squared term of $MathHours_{itk}$ is also included to capture decreasing returns in a simple way. All specifications are augmented with wave-year fixed effects δ_{tk}. The identification strategy is that the variation in $MathHours_{itk}$ is exogenous and provided by the randomized experiment implemented at two different point in time.

Table I1 shows the results below. Column 1 shows the results for the simple regression of math scores onto hours of studying mathematics. One additional hour of studying leads to a higher math score by 7.54 points, or (7.54/13.77=) 0.55 standard deviations. In Table I2 in Appendix I, we add controls in a similar fashion to those in Table 3; that is, we respectively control for the total baseline grade (without math), student characteristics, and all baseline index surveys, and find very similar results.

This function represents a production function of grades, estimated through a randomized intervention implemented at two different time periods: when schools are open and when schools are closed.

In fact, these results allow us to quantify the learning loss. In regular times, the math grade is $7.54 * 3 - 1.16 * 3^2$, whereas during the pandemic, the math grade is $7.54 * 0 - 1.16 * 0$, therefore the learning loss is the difference between these two numbers: 12.18.

This estimate is very close to the standard difference-in-difference estimator (we had found -11.95 for the coefficient of $SchoolClosed$ in Table 3), yet it does not rely on the parallel trends assumption.

Table I1—Hours Worked

	(1)
	Math grade
Math Hours	7.54***
	(1.43)
Math Hours Squared	-1.16***
	(0.33)
Wave*Year fixed effects	YES
Controls:	
Baseline Total grade	NO
Age, Gender, School Year	NO
Baseline Survey	NO
Observations	2,170
R-squared	0.355
Mean Dep. Var.	40.82
SD	13.77

Instead, our estimator relies on a randomized experiment, implemented at two different points in time, such that we can evaluate the decreasing returns to hours of teaching in math in a production function of grades. The fact that these two methodologies yield relatively similar estimates support the claim that school closures causally created a large learning loss.

We corroborate the evidence that we provided earlier for the diminishing marginal returns of hours studying on math grades in Figure I1 as well as Table I1 by adding additional controls. More specifically, we augment the specified model with baseline total grade, age, gender, school year, and baseline survey responses. In all specifications, the number of hours spent studying math is statistically significant and positive, while the hours squared term is statistically significant and negative, albeit with a coefficient of much lower magnitude. This confirms the trend of diminishing marginal returns highlighted in Figure I1.

Table I2—Hours Worked

	(1)	(2)	(3)
	Dependent Variable: Math Grade		
Hours	6.93***	7.14***	7.20***
	(1.45)	(1.50)	(2.10)
Hours Squared	-1.09***	-1.09***	-1.10***
	(0.31)	(0.31)	(0.37)
Wave*Year fixed effects	YES	YES	YES
Controls:			
Baseline Total Grade	YES	YES	YES
Age, Gender, School Year	NO	YES	YES
Baseline Survey	NO	NO	YES
Observations	2,170	2,170	2,170
R-squared	0.393	0.399	0.431
Mean Dep. Var.	40.82	40.82	40.82
SD	13.77	13.77	13.77

Note: Standard errors clustered at the student level in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Column 1 shows the results while controlling for the baseline total grade, with a student's math grade as the dependent variable. Column 2 augments this specification by controlling for a student's age, gender, and the year of schooling they are currently completing. Column 3 adds to column 2 by including in the list of controls the baseline averages of various indices from the survey data. These indices include: oral comprehension, computer proficiency, cross-culture communication, motivation, self-esteem, future aspirations, liking school, liking classes, thoughts about Canada, and thoughts about Kenya.

Dear editor,

Thank you so much for the great comments made by you and the referees, and the opportunity to revise our study.

Please read the pdf version of the paper and not the Word document. The editing team insisted on us preparing a fully editable version on Word, however it does not look good since we had originally written the paper in lyx and compiled it to a pdf (which is not acceptable for the editing team). We ask you to please consider the nice version of the paper in pdf (which I uploaded as the "cover letter" since the system did not allow me to update a pdf as the manuscript). We hope this does not cause any confusion, and we apologize for it.

On the substance of things, it took us much longer than anticipated to finalize our revision since we really wanted to take our time and do a good job addressing all the comments made by the referees.

In particular, we:

- Implemented the psychometric tests that R1 wanted us to implement. This was a great idea, which was definitely missing from the previous version. Overall, we find that the psychometric scales used in this paper have internal reliability, test-retest reliability, convergent validity, divergent validity and predictive validity; which definitely adds strength to the analysis.
- Toned down the policy implications based on your advice
- Rewrote completely the introduction and conclusion based on the referees advice.
- Implemented new statistical tests on peer effects and various robustness checks, as required by the referees
- The entire paper was proofread by a native speaker in English, who brought 66 changes to the paper to make it more formal

More generally, we address in detail all the comments made by the referees (see below our answers in blue). We thank you again for the opportunity to revise the paper. We believe the new version is much stronger.

Please let us know if you have any other comments, we will be very glad to address them.

Regards

Manuscript Number: **EDEV-D-23-01514**

Reviewer #1:

This is an important study which adds to the evidence based on the effect of tutoring on education outcomes, the effect of school closures on education outcomes and aspirations and the role of tutoring in counteracting the effects of school closures. All of these have important implications for policy and practice in education in Kenya and beyond. However, a number of issues need to be addressed before this article can make a significant contribution to the field.

Thank you so much for the great comments you make below, and the opportunity to revise our study. We address in detail all your comments (see below our answers in blue). Thanks to your comments, we believe the new version is much stronger.

Major Comments

1. Given the space devoted to robustness checks and identification strategy in the paper, it is surprising how little attention is given to measurement issues. There is little discussion of how measures were developed or on the quality of the measures (pages 8 & 9). For measures used in the study - especially outcome measures - present psychometric data to demonstrate that the data are reliable and valid measures of the intended constructs.

Thank you very much for this comment, it has pushed us to undertake a new analysis of the reliability and validity of these measures. We now implement a new battery of psychometric tests to show reliability and validity, following your comment.

Starting with internal reliability, Table G1 below displays in column (1) the Alpha Cronbach test of each psychometric scales. The Alpha of the Oral Comprehension, Cross-Cultural Communication, Computer Proficiency, Liking School, Liking Courses, Academic Motivation, and Self-Esteem are all above 0.6, as per the NIH guidelines.[1]

[1] https://www.ncbi.nlm.nih.gov/books/NBK581902/

TABLE G1—INTERNAL RELIABILITY

	(1)	(2)
	Alpha Chronbach	ICC
Oral Comprehension	0.96	0.48**
Cross-Cultural Communication	0.68	0.28**
Computer Proficiency	0.87	0.34**
Aspirations	0.51	0.33**
Liking School	0.70	0.17**
Liking Courses	0.70	0.21**
Academic Motivation	0.64	0.45**
Self-Esteem	0.71	0.28**
Thoughts on Canada	0.41	0.02**
Thoughts on Kenya	0.40	0.28**

This indicates that the items composing a scale are well correlated with each other; i.e., when participants give a high response for one of the items, they are also likely to provide high responses for the other items. The Aspirations scale has a slightly lower alpha (0.51), not far from the guideline of 0.6. The alpha for the scales Thoughts on Canada and Thoughts on Kenya are 0.41 and 0.40. These scales are less central to the analysis, with only a remote link between tutoring and thoughts on Canada and Kenya; indeed we did not detect any meaningful treatment effect for these scales.

For test-retest reliability, we calculate the correlation between repeated waves for the same students. The intraclass correlation is displayed in Column (2). It is above 0.3 for most scales, and always statistically significant at 5 percent.

Convergent validity states that scales measuring the same concepts should positively correlate with each other. In Table G2, we measure the correlation between Oral Comprehension, Cross-Cultural Communication, Computer Proficiency, Aspirations, Liking School, Liking Courses, Academic Motivation and Self-Esteem. The basic intuition is that these scales should be positively correlated, e.g., students motivated in class should also like courses. Indeed we find a positive correlation between all these scales, as displayed in the table. The only exception is the correlation between aspirations and proficiency with computer, which is negative. One may argue that these two concepts are not obviously connected, therefore a low correlation may be expected. In other words, one can be good at computers and have low aspirations, or vice versa.

TABLE G2—CONVERGENT VALIDITY

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Oral Comp.	Cross-Cultural Communication	Computer Proficiency	Aspirations	Liking School	Liking Courses	Academic Motivation	Self Esteem
Oral Comprehension	1							
Cross-Cultural Communication	0.39**	1						
Computer Proficiency	0.31**	0.44**	1					
Aspirations	0.09**	0.06**	-0.07**	1				
Liking School	0.14**	0.25**	0.20**	0.11**	1			
Liking Courses	0.19**	0.27**	0.29**	0.10**	0.95**	1		
Academic Motivation	0.20**	0.21**	0.30**	0.20**	0.45**	0.50**	1	
Self-Esteem	0.14**	0.16**	0.03	0.11**	0.25**	0.22**	0.38**	1

Divergent validity states that there should be no correlation between measures that should not have a relationship. To test for divergent validity, we focus on the scales Thoughts on Canada and Thoughts on Kenya. These scales ask about perceptions on the countries of Canada and Kenya (e.g., Canada is a great place to live; Kenya is a great place to live). These scales should not be connected with academic motivation or liking school; and indeed they are not, as Table G3 shows.

TABLE G3— DIVERGENT VALIDITY

	(1)	(2)
	Thoughts on Canada	Thoughts on Kenya
Oral communication	0.0666*	-0.1520*
Cross-cultural communication	-0.0084	-0.0623*
Proficiency with computer	0.0011	-0.1810*
Aspirations	0.1235*	0.1108*
Liking School	0.0136	0.0606*
Liking courses	0.0385	-0.0176
Academic Motivation	0.0431	-0.0599
Self Esteem	-0.0078	0.0764*

Finally, predictive validity evaluates how well a scale predicts an outcome. We use grades in school as an outcome in Table G4 below, and and find that indeed the psychometric scales of

Oral communication, Cross-cultural communication, Liking School, Liking courses, Academic Motivation, and Self Esteem are positively correlated with grades in school. The scale Aspirations is positively correlated with grades, very close to being significant. The scale Proficiency with computer is not correlated with grades, which may be expected since these are different skills. One can be proficient with using a computer, sending emails, but this does not necessarily correlate with grades.

TABLE G4—PREDICTIVE VALIDITY

	(1)
	Grade Total
Oral communication	0.2189*
Cross-cultural communication	0.0801*
Proficiency with computer	-0.0059
Aspirations	0.0635
Liking School	0.0906*
Liking courses	0.0982*
Academic Motivation	0.1275*
Self Esteem	0.0703*

We believe these new results show that the psychometric scales used in this paper display internal reliability, test-retest reliability, convergent validity, divergent validity and predictive validity. Please let us know if these tests are to your satisfaction.

We included them in the paper in a new appendix G.

Additionally, some of these scales are standard in the literature and have already been used and validated in other contexts, for example the psychometric tools on liking school from Pell and Jarvis (2001), academic motivations from Muris (2001), and self-esteem from Rosenberg et al. (1995).

Pell, Tony, and Tina Jarvis. 2001. “Developing attitude to science scales for use with children of ages from five to eleven years.” International Journal of Science Education, 23(8): 847–862.

Muris, Peter. 2001. “A Brief Questionnaire for Measuring Self-Efficacy in Youths.” Journal of Psychopathology and Behavioral Assessment, 23: 145–149.

Rosenberg, Morris, Carmi Schooler, Carrie Schoenbach, and Florence Rosenberg. 1995. “Global Self-Esteem and Specific Self-Esteem: Different Concepts, Different Outcomes.” Ameri- can Sociological Review, 60(1): 141–156.

The oral proficiency scale is more unique to this paper, we provide a full explanation in Appendix A. The other scales are all explained in Appendix F.

We thank you for this comment which greatly improves the paper, since the results of the new psychometric tests on reliability and validity are largely positive.

2. This is little detail in the introduction on what is meant by 'tutoring'

Thank you for giving us the opportunity to clarify our explanations. We now include the definition of Nickow et al. (2020) right in the first paragraph of the introduction:

“Tutoring - **defined in Nickow et al. (2020) as one-on-one or small-group instructional programming by teachers, paraprofessionals, volunteers, or parents** - might be a valuable option: it causally improves grades…”

In our context, it means the following system (paragraph 3 of the introduction: “In this paper, we implement a randomized experiment on online tutoring in remote rural areas of Kenya. The tutors are university students volunteers. They communicate through the internet on an electronic tablet with their tutees. The tutees are primary school students in rural Kenya. The tutoring was in English for the years 2016 to 2018 and in Maths for the years 2019 to 2020.”

We hope these clarifications are to your liking. Please let us know if you want us to clarify this further.

3. The literature review involves documenting empirical findings from similar studies but there is little attempt to engage with the policy debates in the field. Some questions that should be addressed in the literature review and conclusions:

* What is the function of tutoring in a Kenya education system and in other similar contexts? Who is advocating for it? What problem is it trying to address? How does this study inform these questions?

There was a very active policy debate in Kenya at the time of covid and when the schools were closed about the efficacy of the online distance learning measures put in place by the government to counter the learning loss.

The Kenya Institute for Curriculum Development and UNICEF provided pre-primary and primary lessons, through TV, radio, and internet uploads. Students could access the official education extension material, available on the Kenya Education Cloud (KEC) (see https://kec.ac.ke/). These programs have been widely criticized in Kenya [Ochieng2022, Malenya2023, Mabeya2020]. These papers suggest that few children were able to access these education extension efforts. For children able to access them, the remote lessons moved too quickly for them, were not at the right level, and did not explain the material or solutions in a manner they found accessible.

It is in this context that we suggest the possibility of tutoring as an alternative. Tutoring can alleviate the concerns raised above: tutoring can be personalized at the right level, and it can reach even the rural underserved communities. Yet there was no study demonstrating rigorously the effects of online tutoring. Our paper is the first to do so. The policy implication of our paper is that tutoring can work as an alternative, especially when schools are closed.

We now add all this discussion in the new conclusion of the paper, based on your comment 4.

* What do we know about the effect of school closures on academic achievement? How does this study add to that?

The effect of school closures has been the subject of an intense academic and policy debates, with estimates ranging from 0 to 0.7 SD, the higher estimates being found in remote rural areas (Singh, Romero and Muralidharan (2022); Moscoviz and Evans (2022); Patrinos, Vegas and Carter-Rau (2022); Engzell, Frey and Verhagen (2021); Maldonado and De Witte (2020); Kuhfeld et al. (2020); Azevedo et al. (2020); Hevia et al. (2022)).
The fact that we collected data before and after the pandemic allows us to quantify the learning loss in our context. We compare the evolution in scores of the 2020 cohort to the 2019 one (in the control groups). We find a 0.8 SD reduction in test scores, on the high end of the

estimates provided in the literature, which is consistent with the local context of a remote rural area of a developing country with few alternative online options available.

We now add all this discussion in the new conclusion of the paper, based on your comment 4.

* What strategies are currently being discussed as ways to address the problem of school closures

Thank you for raising this point, which allowed us to better delineate our contribution.

Other strategies than online tutoring are currently being discussed to mitigate the learning loss of closing schools: SMS and 5-10 minute phone calls in Botswana (Angrist, Bergman and Matsheng, 2022), a similar program in Nepal (Radhakrishnan et al., 2021), 30 minute phone calls by teachers in Bangladesh (Beam, Mukherjee and Navarro-Sola, 2022), 30 minute phone tutoring in Bangladesh (Hassan et al., 2022), teacher-student 15-minute mini-tutoring sessions in Kenya (Schueler and Rodriguez-Segura, 2021), and weekly phone tutorials from teachers in Sierra Leone (Crawfurd et al., 2022).

The contribution of our paper is to study for the first time online video tutoring.

We now add all this discussion in the new conclusion of the paper, based on your comment 4.

* What do we already know about the interplay between school attendance / achievement and aspirations? How does this study add to that?

This is a great topic, unfortunately our study does not contribute to this debate since we do not have a randomized intervention on these factors to be able to uncover causal effects. Our only intervention is the tutoring, and we can only analyze the effect of that intervention on various outcomes. It would be fascinating though to study what you mention for sure, we are just not equipped to do so. Please let us know if you see a way we could contribute to this question with the available experiment, or any statistical test you would like us to implement.

4. The conclusion is very brief. Consideration of the above questions will help strengthen it. Also, some comments made throughout the paper could usefully be moved to the conclusion section, to improve the structure of the paper.

Thank you so much for this: we found your questions above so inspiring that we completely rewrote the conclusion to write about your points raised above. We follow exactly your order above to rewrite a new conclusion, please let us know what you think about this.

We believe it makes for a much better new conclusions, highlight the contributions of the paper.

5. Overall, the writing style is informal and imprecise in places, e.g.: "which has nothing to do with the outcome studied" and "all these results pertain to the math intervention"

Thank you so much for raising these points. We agree it was a very informal way of writing.

We now write: “This generates a variation in the number of treated students per classroom **which is independent from** the outcome studied”

The next sentence was also very informal:

“All these results pertain to the Math intervention on the Math grades. Another important finding from Table tab:Math-Grades:2016-18_2020 is from the coefficient English…”

We replace it with the more professional sentence:

“The results above focus on the effect of the Math intervention on the Math grades. We now turn to the English intervention. The first finding from Table tab:Math-Grades:2016-18_2020 relates to the coefficient English …”.

More generally, the entire paper was proof read by a native speaker in English, who brought 66 changes to the paper to make it more formal.

6. The abstract is lacking details - the basics of what the online tutoring program involved, the grade level of students, region of the study, research design, what is meant by 'grade' as the outcome measure - are these school exams? In what subjects?

Thank you so much for suggesting all these points. We have now written a new abstract to address all your questions. Please see below the old and new abstracts.

Old Abstract

We evaluate the effects of an online tutoring program that started in 2016 and continued during the pandemic despite the schools being closed for 9 months in Kenya. We find no effect when the schools are open, but a large effect when the schools are closed (0.4 SD increase in grade in the treatment group versus control group). Since we have data from before the pandemic, we are able to quantify the learning loss due to covid: 0.8 SD. We conclude that online tutoring compensates half of the learning loss.

New Abstract

We evaluate the effects of an online tutoring program that started in 2016 and continued during the pandemic despite the schools being closed for 9 months in Kenya. **Volunteer university students tutored by videoconferences primary school students in Kenya, on the topics of Maths and English. We implement a randomized experiment to test the effects.** We find no effect when the schools are open, but a large effect when the schools are closed (0.4 SD increase in **exam scores** in the treatment group versus control group). Since we have data from before the pandemic, we are able to quantify the learning loss due to covid: 0.8 SD. We conclude that online tutoring compensates half of the learning loss.

Minor Comments

Page 2
Para 1

"The quality of education is poor in low-income countries." This statement needs qualifying. What does "poor quality of education" mean? Is this statement universally true across all low-income countries?

Yes we completely agree, we had struggled with this first sentence. We now erased it, based on the advice of referee 2, and replaced it with its corresponding footnote.

Instead of

“The quality of education is poor in low-income countries”

We now have a more factual statement:

Two thirds of children do not achieve a minimum proficiency level in reading and mathematics in Grade 2[2] despite the ambitions of Sustainable Development Goal 4 for “inclusive and equitable quality education and lifelong opportunities for all.”

Please let us know what you think, but we prefer it since it is a more factual statement than what we had before.

A basic definition of tutoring is needed. Does it mean tuition outside school? One-to-one? Etc

Yes thank you for suggesting this. We now include the definition of Nickow et al. (2020) right in the first paragraph of the introduction:

Tutoring - **defined in Nickow et al. (2020) as one-on-one or small-group instructional programming by teachers, paraprofessionals, volunteers, or parents** - might be a valuable option: it causally improves grades…

Final para - I didn't find the results surprising. I suggest removing 'perhaps surprisingly' and let readers decide if they are surprised.

Yes you are right, we removed this.

Page 3 para 3

"0.8 SD reduction in test scores" - please include a description of the kind of test scores you are referring to, for example "education achievement test scores." The phrase "test scores" is generally understood by economists to mean "educational achievement test scores" but spelling this out resolves any ambiguity (given that other kinds of test scores exist) and makes the paper accessible to other disciplines.

Thank you for this clarification, we now use the phrase "0.8 SD reduction in education achievement test scores"

---

[2] Data from world development indicators.[3] Data from world development indicators.

The summary of the paper (pages 2-4) seems quite detailed to me. It goes into a lot of detail in issues that aren't essential for a summary - robustness checks, comparable studies - and skips some essential details such as the definition of tutoring. Can it be shortened?

Yes, we agree. We rewrote the introduction along your recommendations. We took out entirely 2 paragraphs on robustness checks and comparable studies. The introduction is now shorter than before. Please let us know if this is to your liking, we can further cut out other paragrphs if you think it is appropriate.

Page 5, para 4

"very cheap" - this is subjective. I can imagine the cumulative cost of data connection being prohibitive for some in Kenya. I would just report the data (as you do) and let others decide whether this is cheap, perhaps by including details of average incomes in Kenya.

Yes we agree with you, we removed this mention.

Page 6

Please expand on what you mean by 'grades' so that this can be understood by an international audience. This term is not understood universally. Does this involve an assessment of routine class work? By the teacher? Or the result of an exam?

Thank you so much. We think the clearest way to explain. This is to refer to the age of the students. Grade 6 means 6 years after the start of school at age 6, therefore 12 year old students.

We now write:

“Grade 6 students, i.e., 12 year old students”.

Thank you for mentioning this, you are right, this can be understood better by an international audience.

Page 6 "reassuringly, our results .." consider moving this to the discussion section.

Thank you for this suggestion, we removed that paragraph here, which we agree was out of place.

I had difficulty understanding the source of the data presented in Table 1. Please describe the data collection methods before you present the findings.

Yes you are right. We changed the ordering of the text, to present the findings after explaining how we collected the data. Additionally, we introduce a new distinction in Table 1, explaining that the top two variables are from the administrative data on grades, while the bottom measures come from the surveys. We hope this clarifies the presentation of the descriptive statistics.

Page 9
Experimental design

I suggest starting this section by noting that randomisation was conducted at the individual student level and note the total number of classrooms/schools involved in the process.

Thank you for this comment, we now add: “We randomized half of the grade 6 students **at the individual level…”.**
We had a target number of tutors of 25 per semester. We randomized into one or two classrooms every year for 6 years.

Peer effects - I couldn't follow the reasons why some schools have 25 students and others have fewer (this could benefit from more explanation) but it seems that the variation in number of students receiving the intervention could be related to school characteristics - this is not a major issue since you find no peer effects, but could be acknowledged.

We apologize for our unclear explanations. We believe the best way to explain is with a simple numerical example.

Suppose we have 60 students in Grade 6. Our target number is 25 students treated, so we randomly draw 25 students to be treated, and 35 are control. In the final analysis, we then compare the 25 treated grade 6 students to the 35 control grade 6 students.

In another year, suppose we only have 30 students in Grade 6. This can happen for reasons exogenous to the intervention, i.e., the cohort size shrinks in a particular year. We randomize half of Grade 6 into treatment, such that 30/2=15 students are treated. Our target number is 25 students treated, such that 25-15=10 students still need to be treated. We thus consider Grade 5 students. Suppose there are 40 students in Grade 5. We randomly draw 10 students to be treated, and 30 are the control group. In the final analysis we compare the 15 treated grade 6 students to the 15 control grade 6 students, and the 10 treated grade 5 students to the 30 control grade 5 students, to control for the grade level. In practice, we implement this by having a dummy for grade 5 students, such that students of the same grade are compared with each other.

When the 40 grade 5 students graduate to grade 6, we excluded the 10 students already treated in Grade 5 from the sample. Otherwise, some of these 10 students may have received 2 years of treatment, which would have complicated the analysis since we would then have to differentiate between one year of treatment and two years of treatment. We thus excluded these 10 students from the study, and only considered the 40-10=30 other students as part of the study. These 30 students were the control group in Grade 5 and have thus not received any treatment. We then followed the same procedure, i.e., randomized half of that into treatment, compared the 15 treated to 15 control (excluding the 10 already treated in Grade 5) and complemented with the Grade 5 students to reach our target number of 25 students treated.

This example makes it clear that the intensity of the treatment varies in each classroom every year, which allows us to identify peer effects.

We now include these new explanations in the paper. Thank you very much for making this comment, this has really pushed us to explain more the experimental design.

Also - would a better indicator of peer effects be the *proportion* of students receiving the intervention, rather than the *number*?

Yes you are right, we now include this test in the paper. The results are the same in Column (3) of Table 4 if we consider the proportion of students receiving the intervention rather than the number.

We want to thank you again for the comments you made. We tried to address them all, hopefully in a satisfactory manner. Please let us know if you have any other comments and we will be delighted to address them.

Reviewer #2: I read the paper "Online Tutoring Reduces by Half the Learning Loss due to School Closures: Evidence from a Randomized Experiment in Kenya. This paper uses a randomized experiment to evaluate how students in Kenya respond to both online tutoring and school closures caused by the pandemic. The authors collect novel data, and a strength of the paper is that the experiment and data collection began in 2016, before the pandemic. The main outcome of interest is test scores (and English proficiency), but the authors have included many other measures such as aspirations and technology knowledge.

Thank you so much for the great comments you make below, and the opportunity to revise our study. We address in detail all your comments (see below our answers in blue). Thanks to your comments, we believe the new version is much stronger.

This paper contributes to the literature on online tutoring in developing countries, as well as the impacts of online tutoring during the pandemic. Although not emphasized to the same extent, I also found the results of the intervention on aspirations and computer skills to be of interest, and a notable contribution. I have two main concerns, which I will detail below:

In my opinion, the main results of the paper are the weakest. Looking at Figure 1, it becomes clear that the math test scores are very noisy, and the patterns are very different each year. Given that a lot of the data included in the control is coming from the earlier years (2016-2018) this is worrisome. The sample is very small each year (25 treated students) so to get power all years are combined. I would have no issue with this if the outcome of interest was a variable like "tutoring", however given that they are trying to identify the effects of math tutoring specifically, it is not clear to me the benefit of including the years 2016-2018, given that there was an English tutoring intervention and the math test score patterns look drastically different.

Thank you very much for suggesting this point. It is true that the tutoring was in English for the years 2016-2018 and in Maths for the years 2019-2020. Therefore, you point out that “it is not clear to me the benefit of including the years 2016-2018”.

We performed a new statistical analysis to explore this point. We removed entirely the years 2016-2018 from the analysis. The results are exactly the same as before. Therefore, the inclusion or exclusion of the 2016-2018 years is not driving the result. We now present these new results in Appendix E.

We include the results with the entire sample in the main body of the paper since the results are not different, the sample is larger, and there is a full set of year fixed effects for the period 2016-2018 years, and an “English tutoring” variable different from the Math Tutoring intervention; such that these two periods are essentially analyzed separately.

Continuing this point, there is a large noticeable drop in math scores in the 2020 year between wave 0 and wave 1. In fact, comparing the change between wave 1 and wave 5, 2020 actually has less of a decrease than 2019. This should at the very least be addressed in the paper. It would be helpful to have more information on the tests here - for example, are the tests in each wave identical each year? If so, is there any possible explanation for the different patterns across years? One solution would be to look at the value-added compared to wave 0, however this would require dropping everyone who does not have baseline test scores.

Yes we agree with you that there is a large noticeable drop in math scores in the 2020 year between wave 0 and wave 1. However, the important point is that there is the same noticeable drop in the treatment group and control group. Therefore, when we compare the treatment group and control group, there is no noticeable differential evolution between the treatment and control groups (there is no "difference-in-differences"); which is logical since there is no intervention at that time.

The large noticeable drop between wave 0 and wave 1 that you noticed may be coming from the fact that exams are different across time. The level of difficulty of a test may differ over time. Therefore, one cannot simply look at the evolution of the treatment group over time: this confounds the effect of the intervention and the effect of varying difficulty of the exam. One must compare the treatment group to a control group. When we do so between wave 1 and wave 0, we see no noticeable "difference-in-differences".

Thank you very much for noticing this, we now include it in our discussion, which is a great way to explain the importance of our control group. We now say:

"The main result comes from 2020 (in blue). There is a large noticeable drop in math scores in the 2020 year between wave 0 and wave 1; however the drop is similar in the treatment group and control group. This large drop in both groups may be coming from the varying difficulty of exams. This highlights the importance of having a control group, to control for the difficulty of the exam.
A difference between the treatment and control groups emerges in later waves. Schools were closed for waves 2 through 4 in 2020, hence the missing grades, but online tutoring continued. In wave 5 when the schools reopen and grades are taken again, the treatment group is above the control group by a noticeable 5 percentage points difference."

Thank you very much for suggesting this, this is a great point. We implemented your test, and found exactly the same results, if anything larger: a significant 7.5 coefficient of "Math*School Closed" (as opposed to 5.67 in our main result). We agree with you that it drops everyone who

does not have a baseline scores, therefore we keep the main result in the paper, we add a footnote about this test in our paper.

Some of the treated students were in Grade 5 due to the intervention design. I have several concerns about this, and I feel like there should have been more information given about this subsample. For example, as I understand only treated students are coming from Grade 5, not any control students. So in footnote 12, when it says 27% of students are from grade 5, is this of the whole sample or just of the treated group? Further, are the grade 5 students covering the same material, taking the same tests, and using the same textbooks? If the curriculum follows a different pattern in grade 5 compared to grade 6, then different tests could be more or less challenging at different times, and comparing these grade 5 students in the treated group to only grade 6 students in the control group is a concern. Also, the students who were treated in grade 5 were not treated again in grade 6, but were they part of the control group the next year?

We apologize for our unclear explanations. We believe the best way to explain is with a simple numerical example.

Suppose we have 60 students in Grade 6. Our target number is 25 students treated, so we randomly draw 25 students to be treated, and 35 are control. In the final analysis, we then compare the 25 treated grade 6 students to the 35 control grade 6 students.

In another year, suppose we only have 30 students in Grade 6. This can happen for reasons exogenous to the intervention, i.e., the cohort size shrinks in a particular year. We randomize half of Grade 6 into treatment, such that 30/2=15 students are treated. Our target number is 25 students treated, such that 25-15=10 students still need to be treated. We thus consider Grade 5 students. Suppose there are 40 students in Grade 5. We randomly draw 10 students to be treated, and 30 are the control group. In the final analysis we compare the 15 treated grade 6 students to the 15 control grade 6 students, and the 10 treated grade 5 students to the 30 control grade 5 students, to control for the grade level. In practice, we implement this by having a dummy for grade 5 students, such that students of the same grade are compared with each other.

(This example makes it clear that the intensity of the treatment varies in each classroom every year, which allows us to identify peer effects).

When the 40 grade 5 students graduate to grade 6, we excluded the 10 students already treated in Grade 5 from the sample. Otherwise, some of these 10 students may have received 2 years of treatment, which would have complicated the analysis since we would then have to differentiate between one year of treatment and two years of treatment. We thus excluded these 10 students from the study, and only considered the 40-10=30 other students as part of

the study. These 30 students were the control group in Grade 5 and have thus not received any treatment. We then followed the same procedure, i.e., randomized half of that into treatment, compared the 15 treated to 15 control (excluding the 10 already treated in Grade 5) and complemented with the Grade 5 students to reach our target number of 25 students treated.

We now include these new explanations in the paper. Thank you very much for making this comment, this has really pushed us to explain more the experimental design.

We believe this new explanation can answer all your questions:

-"For example, as I understand only treated students are coming from Grade 5, not any control students"

No, we have a control group of grade 5 students, as explained in the numerical example. This is our fault, we did not explain this well earlier.

-"So in footnote 12, when it says 27% of students are from grade 5, is this of the whole sample or just of the treated group?"

It is of the whole sample. We added the word: "27% **of the whole sample** comes from grade 5"

-Further, are the grade 5 students covering the same material, taking the same tests, and using the same textbooks? If the curriculum follows a different pattern in grade 5 compared to grade 6, then different tests could be more or less challenging at different times, and comparing these grade 5 students in the treated group to only grade 6 students in the control group is a concern.

Yes we agree, but we have a control group of grade 5 students, as the numerical example illustrates.

Also, the students who were treated in grade 5 were not treated again in grade 6, but were they part of the control group the next year?

No the treated in Grade 5 were completely excluded from the study in Grade 6, they were not used in the control group.

Thank you, this last point is great, this was not explained properly in the previous version. We believe the numerical example makes it clear, because we now say:

“we compare the 15 treated to 15 control (excluding the 10 already treated in Grade 5)”

Some smaller comments:
* The first sentence of the paper is very blunt. I would recommend softening this a bit, even just by adding in the words "on average".

Yes we completely agree, we had struggled with this first sentence. We now erased it, based on the advice of referee 2, and replaced it with its corresponding footnote.

Instead of

“The quality of education is poor in low-income countries”

We now have a more factual statement:

Two thirds of children do not achieve a minimum proficiency level in reading and mathematics in Grade 2[3] despite the ambitions of Sustainable Development Goal 4 for “inclusive and equitable quality education and lifelong opportunities for all.”

Please let us know what you think, but we prefer it since it is a more factual statement than what we had before.

* There are several details that were not included in the intro that I would have liked to know. They are discussed later in the paper, but I believe should be moved earlier. They are:
o The tutors come from a Canadian university.

This is now in the abstract

o Are all of the students coming from one school in Kenya?

Yes, this is now in the abstract

[3] Data from world development indicators.

o The students are in grade 6.

This is also in the abstract

* It would be nice to have more information on the tests. For example, are these tests designed by the local school? Or are they standardized tests? Are they each testing different material? Are the tests getting harder - is this why the trend is for them to decrease as the year goes on?

The tests are to the sub county level (schools within the sub county sit for similar tests).
Only grade 8 sit for national wide tests to graduate to secondary school.
We don't have any solid evidence that the tests are getting harder. It is true that the red and blue lines (2019-2020) are below the black line (2016-2018); however the gap is very small towards the end of the year. Therefore it is difficult to conclude that the tests are getting harder.
We added the information on the tests designed at the sub-county level in the paper.

* How was the takeup of the treatment? Did all of the students offered tutoring attend the sessions?

Yes all students offered the tutoring attended most of the sessions because the tutors made an effort to engage all tutees in the sessions by teaching at the right level, and engaging in personal discussions as well. We added this in the paper, thank you for mentioning it. All the students loved the novelty of the approach and there was no student who dropped out of the study entirely.

* From the specification provided, I don't believe that it is possible to disentangle the peer effects from class size effects, given that the number of treated students is highly correlated with class size. Given that there are no significant effects, this is not a huge issue.

This is a great comment, thank you so much. We try to account for class size by considering the proportion of students receiving the intervention rather than the number, to account for different class sizes.

The results are the same in Column (3) of Table 4 if we consider the proportion of students receiving the intervention rather than the number, to account for different class sizes.

* There is a control "Wave 5 * 2016-2018" but no control for "Wave 5 * 2019". This should be

justified.

Yes, there are controls for the full set of wave × year fixed effects δ tk . We chose to report the coefficient of Wave 5 * 2016-2018 since it provides an interesting check, as we explain in the paper.

We explain this better now in the paper:

“An interesting check of the difference-in-differences approach is provided by the variable “Wave 5 * 2016-2018” (the full set of wave × year fixed effects is included, we choose to report only this coefficient in the table because it provides an interesting check).”

I enjoyed reading your paper, and I hope that these comments are helpful.

additional comments:

the reviewers were impressed with the novelty of the study but had doubts about somewhat broad conclusions given the samples size and the somewhat "noisy" data. I think the authors can perhaps meet the criticisms, notably by being more modest in their claims of how much can be deduced on a wider scale from such a small and heterogeneous group. They should also, IMHO, themselves be more forthcoming in the discussion about some of the drawbacks of the study.

Thank you so much for this comment. Yes we acknowledge completely these limitation of our study. We acknowledge this in our new conclusion: “A limitation of our study is external validity since our sample is small and the intervention is implemented in rural Kenya.”. We also toned down the paper in several places to address all the criticisms of the two referees. We thank the referees for their thoughtful comments, which have greatly helped us improve the paper.

We want to thank you again for the comments you made. We tried to address them all, hopefully in a satisfactory manner. Please let us know if you have any other comments and we will be delighted to address them.

Highlights

- We evaluate the effects of an online tutoring program in Kenya
- We find no effect on tests scores when the schools are open, but a large effect when the schools are closed (0.4 SD increase in exam scores in the treatment group versus control group)
- Online tutoring compensates half of the learning loss due to school closures during COVID.

Title Page (with Author Details)

Online Tutoring Reduces by Half the Learning Loss Due to School Closures: Evidence from a Randomized Experiment in Kenya

By Matthieu Chemin*, Jeremy Schneider†‡

Draft: September 9, 2024

We evaluate the effects of an online tutoring program that started in 2016 and continued during the pandemic despite the schools being closed for 9 months in Kenya. Using videoconferences, volunteer students from a Canadian university tutored grade 6 students (12 years old) in a rural school in Kenya, on the topics of Maths and English. We implement a randomized experiment to test the effects. We find no effect when the schools are open, but a large effect when the schools are closed (0.4 SD increase in exam scores in the treatment group versus control group). Since we have data from before the pandemic, we are able to quantify the learning loss due to COVID-19: 0.8 SD. We conclude that online tutoring compensates half of the learning loss.

* Department of Economics, McGill University; Cireq, and Cirano. E-mail: matthieu.chemin@mcgill.ca.
† World Bank and Cirano
‡ We wish to thank all the volunteer tutors who dedicated their time and effort for this project. Without them, this project would not have been possible. We also wish to thank the following staff at ELIMU (Evaluation Impact Unit) for their help with data collection and implementation of the project: Kennedy Bundi, Angela Njeri Gitahi, David Gachoki, Dominica Gachuki, Emma Gilman, Adam Aberra, Thomas Kokossou, Momanyi Mokaya, Simon Newman, Diego Albuja Arellano, Hannah Block. Prof. Chemin gratefully acknowledges financial support from SuperProf.com, Seeds of Change, and the Social Sciences and Humanities Research Council (SSHRC).

Online Tutoring Reduces by Half the Learning Loss Due to School Closures: Evidence from a Randomized Experiment in Kenya

Abstract

We evaluate the effects of an online tutoring program that started in 2016 and continued during the pandemic despite the schools being closed for 9 months in Kenya. Using videoconferences, volunteer students from a Canadian university tutored grade 6 students (12 years old) in a rural school in Kenya, on the topics of Maths and English. We implement a randomized experiment to test the effects. We find no effect when the schools are open, but a large effect when the schools are closed (0.4 SD increase in exam scores in the treatment group versus control group). Since we have data from before the pandemic, we are able to quantify the learning loss due to COVID-19: 0.8 SD. We conclude that online tutoring compensates half of the learning loss.

Two thirds of children fail to achieve a minimum proficiency level in reading and mathematics in grade 2[1] despite the ambitions of Sustainable Development Goal 4 for "inclusive and equitable quality education and lifelong opportunities for all." Tutoring - defined in Nickow, Oreopoulos and Quan (2020) as one-on-one or small-group instructional programming by teachers, paraprofessionals, volunteers, or parents - might be a valuable option: it causally improves grades (see Nickow, Oreopoulos and Quan (2020) for a review of the experimental literature), it is the ultimate customization of learning and reduction in class size, it allows for more engagement, rapid feedback, human connection and mentoring, and it bypasses the systemic issues of education systems in developing countries. The problem is how to reach students in remote rural areas of low-income countries, as well as high costs and the limited local supply of tutors.[2]

In this paper, we explore the potential of online tutoring by volunteers to address these issues. The recent improvements in communication technologies have made it possible for a tutor to teach students even in remote rural areas of developing countries. Having volunteers teach online can both drive down costs and expand the set of tutors available. Importantly, it can continue even if schools shut down. Despite the simplicity of the idea, there is no evidence that

[1] Data from world development indicators.
[2] For example, Romero, Chen and Magari (2021) finds that tutoring with local tutors does not improve grades in Kenya.

this would work in a remote rural area context of a developing country, where the efficacy of the treatment might be negatively affected by the cultural divide between tutors and tutees.
In this paper, we implement a randomized experiment on online tutoring in remote rural areas of Kenya. The tutors are university student volunteers. They communicate through the internet on an electronic tablet with their tutees. The tutees are primary school students in rural Kenya, at the grade 6 level (12 years old). The tutoring subject was English for the years 2016 to 2018, and Maths for the years 2019 to 2020.
A unique feature of our program is that it started in 2016 and because of its online nature, continued uninterruptedly after March 2020 when the schools closed in Kenya for 9 months. The Kenyan Government took time to respond by providing lessons through TV, radio, and the internet. These programs were widely criticized for being inaccessible, difficult to follow, and not adapted to the level of students in rural remote areas, further aggravating inequalities.[3]
In contrast, the tutoring continued uninterrupted. We are thus able to evaluate the effects of the same program at two different points in time, when the schools are open and when they are closed.
We find little effects of this online tutoring program when the schools are open, and a large effect when the schools are closed. When the schools are open, the English tutoring has a modest effect on reading comprehension, and the Maths tutoring has no effect. The results are very different when the schools are closed: we find a large effect on grades in that time period (0.4 SD in Maths, the discipline taught at that time). Thus, online tutoring appears especially effective when no other schooling options are available. Our explanation is decreasing returns to hours of teaching. When the schools are open, the tutoring program (1 hour of Maths per week) comes after a full teaching load (3 hours of Maths per week). We find little effects there. When the schools are closed, the online tutoring is the only source of education (barring the official TV/radio program). We find a large effect at that time.
Online tutoring appears to be critical in the period of school closures due to COVID-19. We dig deeper into this result by first quantifying the learning loss due to school closures, a subject of intense academic and policy debates, with estimates ranging from 0 to 0.7 SD, the higher estimates being found in remote rural areas.[4]
The fact that we collected data before and after the pandemic allows us to quantify the learning loss in our context. We compare the evolution in scores of the 2020 cohort to the 2019 one (in the control groups). We find a 0.8 SD reduction in education achievement test scores, on the high end of the estimates provided in the literature, which is consistent with the local context of a remote rural area of a developing country with few alternative online options available. We conclude that the online tutoring program compensates for (0.4/0.8=) half of the learning loss.
The final finding concerns aspirations. We document a large loss in aspirations when the schools are closed, especially aspirations to go to university. An explanation is that students know that their chances to go to university have been harmed. The online tutoring program

[3] See for example Patrinos, Vegas and Carter-Rau (2022); Olanrewaju et al. (2021); Ochieng and Ngware (2022); Malenya and Ohba (2023); Mabeya (2020).
[4] Singh, Romero and Muralidharan (2022); Moscoviz and Evans (2022); Patrinos, Vegas and Carter-Rau (2022); Engzell, Frey and Verhagen (2021); Maldonado and De Witte (2020); Kuhfeld et al. (2020); Azevedo et al. (2020); Hevia et al. (2022)

does not compensate for this: there is no discernible effect on aspirations in the treatment group compared to the control group.

Overall, we thus conclude that the tutoring program compensates partially (half) for the learning loss on cognitive skills, but does not compensate for the negative effect on aspirations. These results are important for policy implications: while online tutoring holds some promise (at least for cognitive skills), it does not fully substitute for school time. These large learning losses estimated in this paper as well as the large decrease in aspirations must be factored in when deciding on future school closures.

Our paper contributes to a burgeoning literature on tutoring from developed countries (Nickow, Oreopoulos and Quan, 2020; Carlana and La Ferrara, 2021; Kraft et al., 2022). Our study provides the first randomized experiment in developing countries, where education systems have systemic issues and online tutoring has a high potential to reach underserved communities.

In a developing country context, our paper also contributes to a growing literature on ways to mitigate the learning loss of closing schools. Previous studies have found positive results of projects providing SMS and 5-10-minute phone calls in Botswana and Nepal (Angrist, Bergman and Mat- sheng, 2022; Radhakrishnan et al., 2021), 30-minute phone calls by teachers in Bangladesh (Beam, Mukherjee and Navarro-Sola, 2022), 30-minute phone tutoring sessions in Bangladesh (Hassan et al., 2022), but no effects of teacher-student 15-minute mini-tutoring sessions in Kenya (Schueler and Rodriguez-Segura, 2021) or from weekly phone tutorials from teachers in Sierra Leone (Crawfurd et al., 2022). The contribution of our paper is to study online video tutoring. Importantly, our experiment starts prior to the pandemic, in 2016. This allows us to study online tutoring when the schools are open, and also to quantify the learning loss due to the school closures using data from before. We find a large 0.8 SD learning loss. We are thus able to answer the question of how much of the learning loss is mitigated by online tutoring (our answer is half). The scalability of online tutoring depends critically on the supply of college students who are willing to volunteer their time as tutors. The objective of the current paper is more to provide evidence on the likely effects of online tutoring: we find very limited effects of online tutoring when the schools are open, and a larger effect when the schools are closed.

1 Intervention

The intervention consists in offering free tutoring to primary school students living in a rural community of Kenya (Kianyaga, three hours north of Nairobi).[5]

The innovative part of the program is that it is conducted online: the tutoring is done entirely by Skype (and then Zoom). The tutors are Canadian university students who volunteered to become tutors for the program. Students receive one hour of tutoring per week. Tutors and tutees are paired randomly and stay together throughout the school term.

The tutoring was in English for the years 2016-2018 and in Maths for the years 2019-2020. For English tutoring, tutors are trained to undertake “ice-breaking” activities in the first tutoring

[5] See https://elimu.lab.mcgill.ca/pamoja.html for a short video on the program and pictures of the area.

session in order to establish a relationship and to gauge the English knowledge and learning level of tutees.[6]

The tutor then follows the official English textbook and helps tutees with their homework, keeping in mind the actual learning level of the students. The tutors are told that there is no point attempting difficult exercises if the tutee lacks rudimentary skills. Instead, tutors are advised to first build fundamental skills. The tutors are provided a range of techniques to teach at the right level, such as going back to easier exercises, building their own exercises, not following the textbook if they have a better idea or if they think the textbook does not follow a logical order. The emphasis is placed on teaching one simple thing right rather than many complicated ones.

A typical tutoring session in English consists of several minutes of tutors and tutees catching up with each other, followed by the tutee reading the most recent chapter of their English textbook. During the session, the tutor follows along the reading and are encouraged to interject and help their tutee with words that they may find difficult to pronounce and are encouraged to answer the questions of tutees. At the end of each reading, there are questions that both tutor and tutee discuss and cover, to test the tutee's reading comprehension skills on the passage that was just read.

For the math tutoring, tutors are instructed to follow the material currently being taught in the students' math class. Tutors follow the same method of gauging the level of each student, going back in the textbook if they see the students struggling, with the objective to build foundational skills while still following the Kenyan curriculum and the current textbook. Tutors made an effort to engage all tutees in the sessions and all students offered the tutoring attended most of the sessions.

A crucial aspect of the program is that it continued after March 2020 when the schools closed. We deployed tablets in students' homes and offered the data costs to connect to the internet for the single hour of tutoring per week (0.24 USD per one hour session per child). Access to internet was given only for this single hour per week. The tutors continued the exact same tutoring they were providing before. The tutors and tutees made sure to find a calm area. No tutors reported significantly more disturbance than when the tutoring was done in the school. The tutoring sessions were conducted at the exact same time as they would have had the schools been opened.

At that same time, alternative options to schools were offered in Kenya. The Kenya Institute for Curriculum Development and UNICEF provided pre-primary and primary lessons, through TV, radio, and internet uploads. Students could access the official education extension material, available on the Kenya Education Cloud (KEC) (see https://kec.ac.ke/).

Qualitative evidence suggests that few children were able to access these education extension efforts. For children able to access them, the remote lessons moved too quickly for them, were not at the right level, and did not explain the material or solutions in a manner they found accessible. More generally in Kenya, these programs have been widely criticized (Ochieng and Ngware, 2022; Malenya and Ohba, 2023; Mabeya, 2020).

[6] Tutors introduce themselves, and follow a list of questions to ask their tutees (for example, what is your favorite sport/game, movie/TV show, subject at school?). The tutor then asks "what surrounds you?", prompting the tutee to describe the place where he/she is. The tutor also undertakes a "would you rather. . . ?" activity to encourage the tutee to talk about him/herself.

It is in this context that we suggest the possibility of tutoring as an alternative. Tutoring can alleviate the concerns raised above: tutoring can be personalized at the right level, and it can reach even the rural underserved communities. Yet there was no study demonstrating rigorously the effects of online tutoring. Our paper is the first to do so. The policy implication of our paper is that tutoring can work as an alternative, especially when schools are closed.

2 Data

We use administrative data on grades taken 9 times during the year (three per trimester) for grade 6 students, who are typically 12 years old.[7]
We use the last grade in the year before as the baseline grade, to estimate baseline cognitive ability. We thus have one pre-treatment wave and 9 post-treatment waves (T=10).
This large number of repeated waves allows us to have a high statistical power in this study. McKenzie (2012) recommends going beyond the single baseline-single endline paradigm in randomized experiments to include more post-treatment waves, especially if there is low autocorrelation in the outcome studied. In our case, there is a 0.53 autocorrelation in the Maths grades.
The total sample size is 2,439 observations.[8] This sample has enough statistical power to identify a minimum detectable effect size of 2.5 percentage points in grades.[9] Even though our study is statistically powered to detect this effect, the downside of a small N sample is external validity. On key metrics, our sample is representative of the rest of Kenya. Students score on average 41 percent in Math and 231 out of 500 on all fields.[10]
These scores are very similar to national averages.[11]
We complement the administrative data on grades with a survey, collected 4 times per year.[12]

[7] There was an exception made in 2019 and 2020 when the number of grade 6 students was slightly too low and few grade 5 students were entered in the study. The tests are designed at the sub county level (schools within the sub county sit for similar tests).
[8] With one pre-treatment wave and 9 post-treatment waves (T=10) and 299 unique student-year observations, a balanced panel would contain (N*T=) 2990 observations of students' grades. Our panel has fewer observations (2439) for three reasons. First, schools were closed during waves 2, 3, 4 in 2020 due to the pandemic. There was not enough time for the test in wave 7, which is also missing. Second, we were unable to trace grades for grade 5 students in the 8th post-treatment wave of 2019. Finally, there is attrition, with 13 grades missing for the years 2019-2020. We find no differential attrition between the treatment and control groups, as shown in Table 12. Additionally, we implement a test for attrition and find the same results, as described below.
[9] With a significance level of 5%, statistical power of 80%, equal size between treatment and control groups (149 observations each), standard deviation of 14, one pre-treatment wave and 9 post-treatment waves, autocorrelation of 0.53 in the math grade and an ancova method, the minimum detectable effect size is 2.5 percentage points.
[10] Other fields are: English, Swahili, Science, Social Studies, and Religious Studies.
[11] Oketch and Mutisya (2013) report that the proportion of schools scoring 250 marks and above between 2002 and 2011 is 42%. Moreover, disaggregated grades by fields of study are available for Isiolo, a county not far from Kirinyaga county where the study is situated, and the average Math grade is 48 percent, average total grade 241. data available at: https://africaopendata.org/dataset/kcpe-2020-performance-in-isiolo-county
[12] Baseline surveys are conducted at the start of every school year in January, with three follow-up surveys at the start and end of the second term in May and August, and an endline survey at the end of the school year in late October.

When the schools were open, we collected the survey in the school. When the schools were closed, we collected the survey in students' home. This was slightly harder than staying on school grounds and waiting for students to come to school, which explains the slightly smaller sample for 2020 (with 37 missing observations). Thus, our total sample size is 1159 observations instead of the theoretical 1196 observations.

The descriptive statistics in these surveys are also very similar to national averages. In this study, students are 11.8 years old on average.[13]

Within this sample, the proportion of females is 44 percent, once again in line with the Kenyan average of 48 percent.[14]

The communities where the program is implemented share common features with other rural communities in the Central province of Kenya in particular, and Kenya in general. For example, the averages of age, gender, and poverty levels are similar to those of other communities in the 2009 Kenya Population and Housing Census; the 2005 Kenya Integrated Household Budget Survey (KIHBS); and the 2008 Kenya Demographic and Health Survey (DHS) (as found in Chemin (2018)). The particular area was selected in 2007 for a study on the effects of access to electricity, a project which has not yet fully materialized. Therefore, this community was not selected for this particular project on online tutoring.

We develop our own measure of oral proficiency in English, explained in greater detail in Appendix A, using the internationally recognized "Common European Framework of Reference for Languages (CEFR)". Table 1 shows that the average oral proficiency score in the baseline of the control group is 3.10 (out of 6), which corresponds to level A2 (basic user) in the CEFR classification.

We also ask questions on cross-cultural communication, on a scale from 1 (least) to 5 (most) how comfortable they would be talking to someone from another country and how much they would worry about what to say to someone from another country (details in Appendix 4.5). This section allows us to track how the intervention affects the student's comfort speaking and interacting with non-Kikuyu individuals. For many of the students, these interactions were their first times meeting someone who comes from outside the local community. Table 1 shows that the average is 3.87 out of 5, this includes the entire sample with the effect of the treatment.

We then ask questions on computer proficiency, explained in detail in Appendix G.0.1. This section is designed to track how the intervention affects the student's computer and technology proficiency over time. For many of these students, this was their first times using a computer, as evidenced by the very low average over these five questions (2.08 out of 5) in Table 1.

We also include in our survey measures on aspirations, related to higher education, career, and broader goals in life. We ask students whether they desire to go to university, their desired age to marry and number of kids, what future career they would like to pursue, and other similar questions. Since questions are on different scale, we standardize all the variables, calculate the unweighted average, and re-standardize on the baseline wave. The purpose for these questions is to see how students may be motivated to continue staying in school. For example, if a

[13] 27% of the whole sample comes from grade 5 since we included few grade 5 students in 2019 and 2020 to increase the sample size, as explained above.

[14] p.291 of the 2021 Economic Survey available at: https://www.knbs.or.ke/wp-content/uploads/2021/09/Economic-Survey-2021.pdf

student says that they would like to marry at a later age, this could indicate that the student wants to carry on with higher education and a career first, similar to their response on how many kids they would like to have. Since we also ask students what their desired future career would be, we can see whether students want to take on jobs that are more human-capital intensive and require higher education, such as lawyers, doctors, nurses, or if they want to take on other vocations which may not require formal schooling such as army or police officers, performers or professional athletes. With the intervention, we expect treated students to want to take on more human-capital intensive careers.

We also include other psychometric tools on liking school from Pell and Jarvis (2001), academic motivations from Muris (2001), self-esteem from Rosenberg et al. (1995), and perceptions of life in Canada and in Kenya to test whether the treatment affects these factors. All of the questions are explained in detail in Appendix G. In Appendix H, we find that the psychometric scales used in this paper display internal reliability, test-retest reliability, convergent validity, divergent validity and predictive validity.

Table 1 shows that the average response on the liking school index is 3.88 out of 5, motivation is 3.25 out of 5, self-esteem is 2.95 out of 5, and perceptions about Canada and Kenya is 0.93 and 0.86 out of 1 (where a value closer to 1 indicates a better perception). Overall, student generally like school, are motivated, and have a good perception of both Canada and Kenya.

Table 1: Descriptive Statistics

	(1)	(2)	(3)
	Mean	SD	Count
Administrative data on test scores:			
Maths	41.3	13.89	2439
Grade Total	231.4	45.99	2435
Surveys:			
Age	11.81	1.16	286
School Year 5	0.27	0.44	290
Female	0.44	0.50	286
English Proficiency	3.06	1.19	1061
Cross-Culture Communication	3.86	0.79	1071
Computer Proficiency	2.08	0.99	1031
Aspirations	-0.27	1.00	1071
Liking School	3.88	0.35	1071
Motivation	3.25	0.52	1071
Self-Esteem	2.95	0.26	1071

Thoughts on Canada	0.93	0.16	1071
Thoughts on Kenya	0.86	0.12	1071

Note: Summary statistics for variables related to students' academic performances and baseline survey responses. Each of the variables after Female represent the baseline averages of an index consisting of various social and psychological questions related to the given topic. Apart from Aspirations, Thoughts on Canada, and Thoughts on Kenya, each of the indices can range from one to five. The aspirations index is standardized due to several of its components having different ranges, and the two indices related to thoughts on Canada and Kenya are comprised of variables that ranged from 0 to 1.

3 Experimental Design

The way we randomized our sample is the following: we had a target number of 25 tutors per semester. We randomized half of the grade 6 students at the individual level into the treatment group. When the total size of the grade 6 class was more than 50 students, we simply selected 25 students from grade 6 to become the treatment group. When the total size was less than 50 students (this happened in the years 2019 and 2020 during the Math treatment), we randomized half of the class into the treatment group. This means less than 25 students are treated. Since our number of tutors was 25, we then consider grade 5 students and pick the rest of the treated students from grade 5. There is thus a treatment group and control group of grade 6 students, and a treatment group and control group of grade 5 students.
When the treated students from grade 5 graduated to grade 6, we faced the choice of selecting them again for treatment in grade 6. This could have generated a treatment of 2 years for some. To keep things simple and limit the intervention to at most 1 year per student, we decided to exclude these treated students from the randomization of the next year. Thus, every treated student has at most received the treatment 1 year. We thus exclude these students treated when they graduate into grade 6, and select the new treated students from the rest of the sample.[15]

[15] A numerical example can be used here to illustrate the experimental design. Suppose first we have 60 students in grade 6. Our target number is 25 students treated, so we randomly draw 25 students to be treated, and 35 are control. In the final analysis, we then compare the 25 treated grade 6 students to the 35 control grade 6 students.
In another year, suppose we only have 30 students in grade 6. This can happen for reasons exogenous to the intervention, i.e., the cohort size shrinks in a particular year. We randomize half of grade 6 into treatment, such that 30/2=15 students are treated. Our target number is 25 students treated, such that 25-15=10 students still need to be treated. We thus consider grade 5 students. Suppose there are 40 students in grade 5. We randomly draw 10 students to be treated, and 30 are the control group. In the final analysis we compare the 15 treated grade 6 students to the 15 control grade 6 students, and the 10 treated grade 5 students to the 30 control grade 5 students, to control for the grade level. In practice, we implement this by having a dummy for grade 5 students, such that students of the same grade are compared with each other.
When the 40 grade 5 students graduate to grade 6, we excluded the 10 students already treated in grade 5 from the sample. Otherwise, some of these 10 students may have received 2 years of treatment, which would have complicated the analysis since we would then have to differentiate between one year of treatment and two years of treatment. We thus excluded these 10 students from the study, and only considered the 40-10=30 other students as part of the study. These 30 students were the control group in grade 5 and have thus not received any treatment. We then followed the same procedure, i.e., randomized half of that into treatment, compared the 15 treated to 15 control

This randomization generates a variation in the number of treated students per classroom which is independent from the outcome studied, and only caused by our randomization process. In some classrooms, the number of treated students is 25. In others, the number of treated students is less, equal to half of the total size of the classroom. In yet other classrooms, the number of treated students is small, equal to the difference between the 25 tutors available and the number of students selected for treatment in grade 6.

We use these variations to identify peer effects. The basic idea of peer effects is that more treated students in a classroom should be associated with a positive effect on the control students. Importantly, the number of treated students in our case is independent from the outcome studied (the math grades) and are only related to the randomization process we used.

Aside from being able to measure peer effects, we argue that the experimental design sheds light on important questions. Recall that the schools closed in 2020. At that point, we distributed the tablets in the students' homes to continue the tutoring. We can compare the treatment effect in 2020 and in 2019, when the schools are closed or open. We are thus able to explore the temporal external validity of the results.

Moreover, we can quantify the learning loss due to the school closures by comparing the evolution of the control group in 2020 before and after schools resumed, compared to the evolution of the control group of the previous cohorts over the same time period.

We can then compare the treatment effect in 2020 to that learning loss to answer the question of how much of the learning loss is recovered by the program.

Table 2 shows the balance test. The important result from this table is that the grades are well balanced between the treatment and control group: the treatment group scores the exact same grade: 44 percent in Math and 195 on other fields of study. The differences are not statistically significant. The treatment group and control groups are thus comparable before the intervention starts.

Table 2: Balance Test: Treatment vs Control Group for Grade 6

	(1)	(2)	(3)	(4)
	Control	Treatment	Control-Treatment	P-value
Math Grade	44.21	44.38	-0.17	(0.94)
Grade Total (No Maths)	194.99	195.17	-0.18	(0.97)
Age	12.08	11.58	0.50*	(0.07)
Gender	0.42	0.49	-0.07	(0.33)
Other Cognitive Skills				
Oral Comprehension	2.91	2.71	0.19	(0.29)
Computer Proficiency	1.40	1.38	0.02	(0.87)

(excluding the 10 already treated in grade 5) and complemented with the grade 5 students to reach our target number of 25 students treated.

Cross-Culture Communication	3.48	3.46	0.02	(0.86)
Non-Cognitive Skills				
Aspirations	0.16	-0.02	0.18	(0.21)
Liking School	3.83	3.80	0.04	(0.34)
Motivation	3.15	3.12	0.04	(0.60)
Self-Esteem	2.94	2.89	0.05	(0.15)
Thoughts on Canada	0.98	0.96	0.02	(0.39)
Thoughts on Kenya	0.88	0.86	0.01	(0.44)

Note: Two-sample t-test results for baseline averages of variables related to students' academic performances and survey responses between treatment and control group. Columns 1 and 2 show the mean of the variable at baseline for the control and treatment groups respectively. Column 3 reports the t-test for the equality of means in the control and treatment groups, and column 4 shows the p-value of that difference. The baseline grades for Maths and Grade Total are taken as the final grades from wave 9 of the previous year.

The average age of the control group is 12 years old, 11.6 for the treatment group. There is a slight imbalance here. It is unclear whether older or younger students should react more or less to the treatment. We control for age in all regressions and find very similar results with or without this control.

Aside from this lone difference, none of the other variables are significantly different between the treatment and control groups.[16]

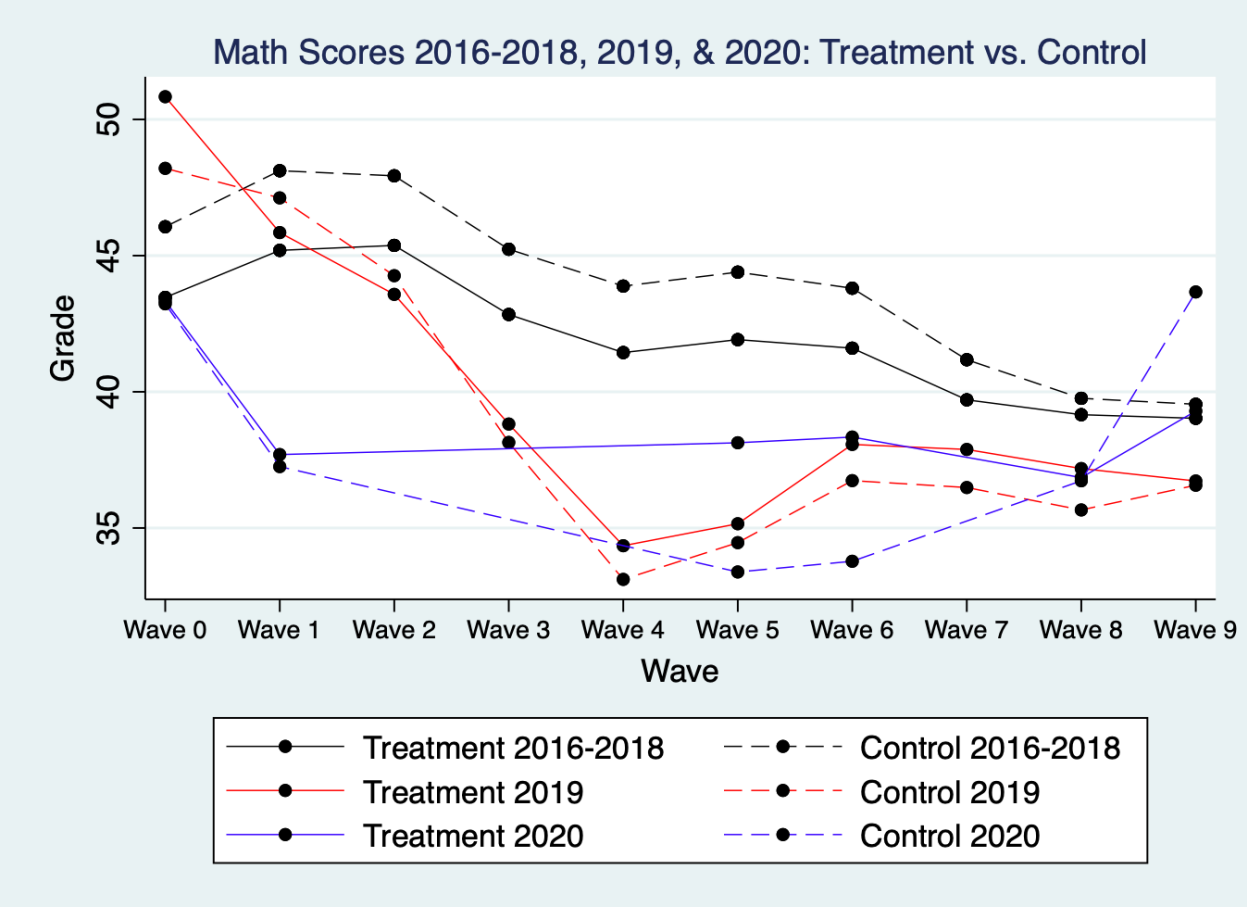
[16] We obtained ethical approval for this study (REB File: 211-1015). There is no pre-analysis plan for this project designed in 2015, however we present in this paper all the outcomes of our questionnaire. We follow the recommendations of Banerjee et al. (2020), and present in the appendix the equivalent of a "populated" PAP, i.e., all the outcomes from the questionnaire. In this paper, we depart from presenting all these outcomes as in a populated "PAP" since we made an important ex-post discovery: we found no effect of the intervention when the schools were open and an effect when the schools were closed. This allows us to estimate a production function of grades featuring decreasing returns, which we use to simulate the effect of closing schools. We had not pre-specified this approach since there was no way of knowing ex-ante that the pandemic would close down the schools for 9 months in March 2020. This new exposition of the results is in line with Banerjee et al. (2020)'s recommendation of "presenting in the paper what was actually learned in the course of the experiment, as opposed to what was anticipated ex-ante".

4 Empirical Analysis

4.1 Effects on Math Grades

We show the raw data on Math grades in Figure 1 below. Wave 0 is the baseline and the treatment is implemented for waves 1 through 9. The black lines show the 2016-2018 period when the intervention was in English. The treatment group is in a solid line, and the control group is in a dashed line. As can be seen on the graph, the treatment has no effect on Math grades, which is logical since the intervention was in English at that time. This is actually reassuring for the integrity of the experiment: the treatment group and control group are on very similar trends absent the treatment (in Maths).

Figure 1: Math Grades: 2016-2018 vs 2020



Note: The figure shows the raw math grades. Wave 0 represents the baseline Math grades (calculated as the last grade of the student in the year before). Waves 1-9 represent the respective periods in the school year. The Kenyan school year begins in January and is divided up into three trimesters, with each trimester containing three periods (and thus nine total periods in a school year). The period 2016-2018 is in back (English tutoring). The treatment group is the solid line, the dashed line is the control group. The year 2019 is in red (Maths tutoring,

schools open). The year 2020 is in blue (Maths tutoring, schools closed). Schools were closed for waves 2 through 5 in 2020, hence the missing grades, but online tutoring continued.

For the year 2019 (in red), we also see no effect of the Math tutoring program. Recall that the schools were open at that time.

The main result comes from 2020 (in blue). There is a large noticeable drop in math scores in the 2020 year between wave 0 and wave 1; however the drop is similar in the treatment group and control group. This large drop in both groups may be coming from the varying difficulty of exams. This highlights the importance of having a control group, to control for the difficulty of the exam.

A difference between the treatment and control groups emerges in later waves. Schools were closed for waves 2 through 4 in 2020, hence the missing grades, but online tutoring continued. In wave 5 when the schools reopen and grades are taken again, the treatment group is above the control group by a noticeable 5 percentage points difference.

The effect disappears over time as the schools reopen. In fact, the control group seems to outperform the treatment group by wave 9, although this effect is not statistically significant (see Figure 2 in Appendix B with confidence intervals). The only significant result in this graph is the positive effect of the treatment when schools were closed.

Thus, we conclude that the math tutoring program has little effect when the schools are open, but has a noticeable impact when the schools are closed. An explanation is that during the pandemic, this online tutoring program was the only source of education (barring the official TV/radio program). Thus, one hour of math tutoring has a large impact when the schools are closed, much less so when the schools are open.

To gauge whether this finding is statistically significant, we use the following specification:

$$y_{itk} = \beta_1 Math_{it} + \beta_2 SchoolClosed_{itk} + \beta_3 Math*SchoolClosed_{itk} + \beta_4 English_{it} + \beta_5 BaselineMathGrade_{it0} + \beta_6 BaselineMissing_{it0} + \beta_7 X_{it} + \beta_8 \delta_{tk} + \varepsilon_{itk} \quad (1)$$

where y_{itk} is the dependent variable in year t, wave k for student i, $Math_{it}$ is a dummy variable equal to 1 if the student was in the Math treatment group (in the years 2019 and 2020), $SchoolClosed_{itk}$ is a dummy variable equal to 1 if year $t = 2020$ and wave $k = 5$, and $Math*SchoolClosed_{itk}$ is the interaction of the Math Treatment dummy and the School Closed dummy. $English_{it}$ is a dummy variable equal to 1 if student i was in the treatment group (in the years 2016 to 2018).

The regression includes the full set of wave × year fixed effects δ_{tk}. Keeping in line with Figure 1, we aggregate the years 2016-2018 together in the wave fixed effects. For example, there is one dummy for wave 5 × Years 2016-2018. We report this coefficient in the main table, to show that there is nothing special about wave 5 in other years. The results are exactly the same if we disaggregate the years 2016 to 2018 in different wave*year fixed effect. Wave fixed effects for the years 2019 and 2020, however, remain separate, since these years are different due to the pandemic. In each column, the wave 1 of 2016-2018 is the omitted wave in the regression.

Some students are missing baseline grades in a given year, but have grades available throughout the school year period. To avoid losing this data in the regressions we run, we use

the following method: for students that have baseline values available, this value is represented in the control variable BaselineMathGrade_it0 . If, however, the baseline value is not available, the value of the control variable BaselineMathGrade_it0 is set to zero and a dummy variable BaselineMissing it is set equal to one. This allows us to keep all of the data available even when the baseline value is missing.

X_it represents a vector of control variables, including age, gender, school year, and students' baseline responses to survey questions related to their levels of motivation, self-esteem, future aspirations, how much they like school in general, and how much they enjoy their classes. Table 3 shows the results of this regression for students' math grades in the school years 2016-2020. Standard errors are clustered at the student level.

Column (1) of Table 3 shows that the variable Mat h it is not statistically different from 0, indicating that being in the educational program did not have any significant effects, at least when the schools are open. The result changes when the schools are closed: the coefficient of Math*SchoolClosed it is statistically significant at the 5% level and has a coefficient of 5.67, indicating that in wave 5 of 2020, being in the treatment group was associated with a math score 5.67 points higher compared to the control group, exactly like in Figure 1.

Table 3: Math Grades: 2016-2018 vs 2020

	(1)	(2)	(3)	(4)
	Dependent Variable: Math Grade			
Math	-0.66	-0.68	-0.52	-0.54
	(1.20)	(1.13)	(1.14)	(1.18)
Math * School Closed	5.80**	6.46**	6.61**	6.33**
	(2.69)	(2.66)	(2.71)	(2.76)
Fisher (p-val)	(0.055)	(0.029)	(0.038)	(0.051)
Wild Cluster Bootstrap (p-val)	(0.11)	(0.11)	(.1)	(0.075)
Attrition: Lower Bound	4.82*	5.29*	5.28*	5.29*
	(2.78)	(2.75)	(2.82)	(2.89)
Attrition: Upper Bound	6.38**	6.92**	7.01**	7.15**
	(2.79)	(2.74)	(2.75)	(2.81)
School Closed	-11.95***	-11.00***	-11.55***	-11.68***
	(2.09)	(2.19)	(2.40)	(3.56)
Wave 5 * 2016-2018	-0.22	-0.11	-0.10	-0.16
	(1.13)	(1.13)	(1.13)	(1.13)
English	-1.09	-0.81	-1.00	-1.06

	(1.39)	(1.32)	(1.30)	(1.20)
Wave*Year fixed effects	YES	YES	YES	YES
Controls:				
Baseline Grade	NO	YES	YES	YES
Age, Gender, School Year	NO	NO	YES	YES
Baseline Survey	NO	NO	NO	YES
Observations	2,170	2,170	2,170	2,170
R-squared	0.355	0.393	0.399	0.431
Mean Dep. Var	40.82	40.82	40.82	40.82
SD	13.77	13.77	13.77	13.77

Note: Standard errors clustered at the student level in parentheses. *** p<0.01, ** p<0.05, * p<0.1. For each column, the dependent variable is a student's math grade in a given wave. 'Math' is a dummy equal to 1 if a student was in the Math treatment group for the years 2019-2020. 'School Closed' is a dichotomous variable equal to 1 when schools are closed, i.e., wave 5 of 2020. 'MathTreatment * School Closed' is the interaction between the two variables. 'Wave 5 * 2016-2018' is a dummy equal to 1 if the period is wave 5 in any of the years 2016 - 2018. 'English' is a dummy variable equal to 1 if a student was in the treatment group for the years 2016 - 2018. All regressions include a full set of interactions between the 9 waves and the 3 time periods (2016-2018 for the English treatment, 2019 for the Math treatment, and 2020 for the Math treatment combined with school closure). All regressions control for baseline math grade and a dummy variable 'Baseline Missing' that is equal to 1 if the baseline math grade is missing. If baseline math grade is missing, it is replaced by the value 0 and the dummy variable 'Baseline Missing' takes the value 1. Column 1 shows the estimation of Equation 1 without any additional controls. Column 2 augments this specification by including the baseline grade on all other topics than Maths, and a dummy variable 'Baseline Missing Total Grade' equal to 1 if the baseline total grade is missing. Column 3 adds to column 2 by controlling for a student's age, gender, and the year of schooling they are currently completing. Finally, column 4 includes in the list of controls the baseline averages of various indices from the survey data. These indices include: oral comprehension, computer proficiency, cross-culture communication, motivation, self-esteem, future aspirations, liking school, liking classes, thoughts about Canada, and thoughts about Kenya.

The variable SchoolClosed (which is simply the dummy for wave 5 of 2020) measures the learning loss, according to the existing difference-in-differences literature, by comparing the evolution of the control group of the 2020 cohort to the control group of the 2016-2018 cohorts. This estimate of learning loss has to be interpreted with caution since it relies on the (untestable) parallel trends assumption, i.e., the 2020 cohort would have evolved the same way as the 2016-2018 cohorts absent the pandemic. Results indicate a decrease in math scores by 12 points. If we interpret this as the learning loss due to the school closures, then this would mean that the educational program alleviates ((5.67/11.95)*100)=47% of the learning loss. This is exactly what is shown in Figure 1, where the difference in math scores between the control group of 2016-2018 and the control group of 2020 is nearly 12 points between wave 5 and wave 0, and only 7 points in the treatment group.

An interesting check of the difference-in-differences approach is provided by the variable "Wave 5 * 2016-2018" (to reiterate, the full set of wave × year fixed effects is included, we choose to report only this coefficient in the table because it provides an interesting check). It is not significantly different from zero, which shows that there are no differences between wave 5 and wave 1 in the years 2016-2018. Thus, if we are willing to make the assumption that there would not be any difference either between wave 1 and wave 5 in the 2020 cohort, then the coefficient of SchoolClosed can be interpreted as the learning loss.

The results remain stable when we control for several variables in the rest of the table. Column (2) adds baseline grades excluding Math as a control variable to the specification in Column (1). The coefficients are largely unchanged. Therefore, controlling for baseline ability doesn't affect the results. Column (3) further builds on the previous specification by additionally controlling for various student characteristics included in the variable (i.e., age, gender, and the school year that the student is completing).

We also include the baseline value of the indices of each of the 10 sections of our survey, namely English oral comprehension, computer proficiency, cross-culture communication, motivation, self-esteem, aspiration, liking school, liking courses, thoughts on Canada, and thoughts on Kenya. The results are exactly the same when we control one by one for each of these indices as in Table 8 in Appendix C; or all of them together in Column (4) of Table 3. We also control for the 51 individual components of these indices, one by one or together, and still find the same effect for Math*SchoolClosed_itk as shown in Table 9. Our results are thus not driven by baseline differences in cognitive or non-cognitive skills between the treatment and control groups.

The results are also the same if we disaggregate the wave fixed effects for the years 2016 to 2018 into different dummies, as can be seen in Table 10 in Appendix D. The results are very similar if we restrict the sample to the years 2019 and 2020 alone when the tutoring was in Maths, as shown in Appendix E Table 11.[17]

4.2 Robustness Checks

We present three robustness checks to adjust the standard errors for the small number of students. First, we use the exact Fisher test (Young, 2018). This permutation test is an exact test regardless of sample size or distribution of error term, as opposed to conventional t-tests which depend on the assumption of large samples (to use asymptotic results), a condition that may be violated in the sample we use, or a normal distribution of the error term. To implement this procedure, we obtain the observed t-stat for the outcome in question, permute the observations randomly between the treatment and control groups, obtain a simulated t-test, repeat this 1,000 times, and find the proportion of occurrences the simulated t-stat is above the observed t-stat, which is the Fisher p-value. In Column (1), the Fisher p-value is 0.055.

Second, we provide a test for the clustering. In our preferred specification, we cluster the standard errors at the level of students. Yet, they could also be clustered at the level of cohorts, which are few (6 cohorts during 4 years). We use the Wild Cluster Bootstrap methodology described in Cameron, Gelbach and Miller (2008) to address this issue. Using Monte Carlo simulations with 6 clusters and different error structures and cluster sizes, they show that

[17] The results are also the same if we look at the value-added compared to wave 0 in our specification.

cluster-robust standard errors reject the null at a rate of 8.2 percent to 18.3 percent. The intuition of the Wild Cluster Bootstrap methodology is to resample residuals at the level of a cluster, thereby preserving the clustering of the data. With 6 clusters, they show that this technique rejects the null at a rate of 1.9 percent to 5.3 percent, not significantly different from 5 percent. In our analysis, we use the 6-point weight distribution proposed by Webb (2014). We find that the results are robust to this correction, especially in the most preferred specification when adding controls in Columns (3) and (4).

Finally, we address the issue of attrition. There is little attrition in the math grades since the data is administrative at the school level (13 missing observations for the years 2019-2020). We find that there is no differential attrition between the treatment and control groups, as shown in Table 12. Moreover, we propose a test for attrition using Manski bounds. We replace the missing observations in the treatment group by the minimum observed value, and in the control group by the maximum value. This represents in a way a worst-case scenario for our estimate. Column (1) of Table 3 shows that the main result is still statistically significant in this worst-case scenario. We also present the best-case scenario, in which we replace the missing observations in the treatment group by the maximum observed value, and in the control group by the minimum value. This builds an upper bound for our estimate. Since the lower bound of the worst-case scenario is still statistically significant, we conclude that the problem of attrition is unlikely to bias our estimates.

4.3 Peer Effects

Recall that because of the way we randomized, there is exogenous variation in treatment intensity: in some classrooms, there were 25 students treated, in others less (if the class size was below 50 students) and in grade 5 classrooms, there were few students treated (to be precise, the difference between 25 and the number of students treated in the grade 6 classroom since our target number of tutors was 25). These variations are exogenous to the outcome studied and solely dependent on our randomization process.

We simply count the number of students treated by the Math intervention per classroom in a variable called “Number Treated Math” and include it in our regressions. More treated students should be associated with a better performance of the control group according to the logic of peer effects. Since the variable “Math” is included, this variable must be interpreted at Math=0, i.e., it represents the increase in Math grades in the control group due to a greater number of students treated by the Math intervention.

Column (1) of Table 4 repeats the main analysis. Column (2) of Table 4 adds this new variable “Number Treated Math”. We find no effect of this variable on the Math grades. In fact, the inclusion of this variable makes no difference to the main coefficient of “Math * School Closed” studied in this paper. More treated students do not lead to a better performance of the control group.

The results are the same in Column (3) if we consider the proportion of students receiving the intervention rather than the number, to account for different class sizes.

Table 4: Peer Effects in the Classroom

	(1)	(2)	(3)

	Dependent Variable: Math Grade		
Math	-0.61	-2.36	0.69
	(1.14)	(2.79)	(2.62)
Math * School Closed	6.33**	6.08**	6.38**
	(2.67)	(2.67)	(2.70)
School Closed	-11.24***	-11.24***	-11.24***
	(2.13)	(2.14)	(2.13)
Wave 5 * 2016-2018	-0.10	-0.10	-0.10
	(1.13)	(1.13)	(1.13)
English	-0.88	-0.88	-0.88
	(1.31)	(1.31)	(1.31)
Number Treated Math		0.14	
		(0.22)	
Proportion treated Math			-3.24
			(6.46)
Observations	2,170	2,170	2,170
R-squared	0.392	0.393	0.392
Mean Dep Var	40.82	40.82	40.82
SD	13.77	13.77	13.77

Note: Standard errors clustered at the student level in parentheses. *** p<0.01, ** p<0.05, * p<0.1. For each column, the dependent variable is a student's math grade in a given wave. 'Math' is a dummy equal to 1 if a student was in the Math treatment group for the years 2019-2020. 'School Closed' is a dichotomous variable equal to 1 when schools are closed, i.e., wave 5 of 2020. 'MathTreatment * School Closed' is the interaction between the two variables. 'Wave 5 * 2016-2018' is a dummy equal to 1 if the period is wave 5 in any of the years 2016 - 2018. 'English' is a dummy variable equal to 1 if a student was in the treatment group for the years 2016 - 2018. All regressions include a full set of interactions between the 9 waves and the 3 time periods (2016-2018 for the English treatment, 2019 for the Math treatment, and 2020 for the Math treatment combined with school closure). All regressions control for baseline math grade and a dummy variable 'Baseline Missing' that is equal to 1 if the baseline math grade is missing. If baseline math grade is missing, it is replaced by the value 0 and the dummy variable 'Baseline Missing' takes the value 1. In Column (2), the variable “Number Treated Math” is the number of students treated by the Math intervention per classroom. In Column (3), the variable “Proportion Treated Math” is the proportion of students treated by the Math intervention in the classroom.

The failure of our statistical test in Table 4 to detect peer effects and the natural absence of peer effects when the schools are closed make it unlikely that our results would be confounded by any peer effects.

4.4 Effects on English Proficiency

The results above focus on the effect of the Math intervention on the Math grades. We now turn to the English intervention. The first finding from Table 3 relates to the coefficient English : the online tutoring intervention in English (organized in 2016-2018) did not increase the math grades. One could have expected a positive impact there since the math textbook is in English, whereas students' home language in this area is Kikuyu, the local dialect. A better mastery of English could have increased the math grades. This is not what we find.

To look more directly at the effects of the English intervention on English proficiency, we use our own assessment tool of oral comprehension in English explained in detail in Appendix A and that follows the in the CEFR classification. This information was collected in our surveys collected 4 times a year (hence the smaller sample size). We build an index of four measures: understanding, conversation, vocabulary, and spoken fluency.

Column (1) of Table 5 shows that the average oral proficiency score in the baseline of the control group is 3.07 (out of 6), which corresponds to level A2 (basic user) in the CEFR classification.

The English tutoring (implemented in the 2016-2018 period) increases this outcome by a statistically significant 0.21 (out of 6). This corresponds to a (0.21/1.215 =) 0.17 standard deviations increase in overall oral comprehension. Thus, online tutoring in English is associated with beneficial effects on English proficiency.

Table 5: Oral Comprehension

	(1)	(2)	(3)	(4)	(5)
	Index	Understanding	Conversation	Vocabulary	Spoken Fluency
Math	0.18	0.15	0.11	0.23*	0.19
	(0.12)	(0.14)	(0.15)	(0.12)	(0.14)
Math * School Closed	-0.38**	-0.33	-0.29	-0.22	-0.25
	(0.17)	(0.21)	(0.21)	(0.20)	(0.23)
School Closed	-1.43***	-1.17***	-1.41***	-0.84***	-1.09***
	(0.24)	(0.26)	(0.24)	(0.24)	(0.25)
English	0.21*	0.21	0.22	0.14	0.20
	(0.13)	(0.14)	(0.13)	(0.14)	(0.14)
Observations	812	813	813	813	812
R-squared	0.491	0.435	0.434	0.370	0.441
Mean Dep. Var	3.074	3.381	3.043	2.943	2.927
SD	1.215	1.254	1.292	1.204	1.351

Note: Standard errors clustered at the student level in parentheses. *** p<0.01, ** p<0.05, * p<0.1. In Column 1, the dependent variable is the unweighted average of the four components. Columns 2 through 5 show the results

of the same regression specification but with each individual component of the oral comprehension index as the dependent variable. The components of this index all range from 1 to 5, with 1 indicating a poor oral comprehension and 5 indicating strong comprehension, and include: understanding, conversation, vocabulary, and spoke fluency. All regressions control for a student's age, gender, and current year of schooling, as well as the baseline survey response of the dependent variable and a dummy variable "Baseline Missing" equal to 1 if the baseline survey response is missing. If baseline survey response is missing, it is replaced by the value 0 and the dummy variable "Baseline Missing" takes the value 1.

We also find that school closures had a detrimental effect on oral comprehension. Just in the period of school closures alone, the average oral comprehension level dropped by more than 1 standard deviation (1.43/1.215). The effect of the school closure was not compensated by the Math tutoring intervention, which is quite logical since the tutoring was in Math.

4.5 Effects on Cross-Cultural Communication

A key question with online tutoring in a cross-cultural context is whether the cultural divide may negatively affect the tutoring.

In fact, we find in Table 6 shows that the treatment leads to overall higher student capabilities in cross-cultural communication. Column (1) shows the unweighted average of our two questions on the topic: "How comfortable would you be talking to somebody from another country?", and "How much would you worry about what to say if you were talking to someone from another country?".

The tutoring in English improves cross-cultural communication comfort by (0.41/0.791) = 0.52 standard deviations compared to the control group. Similarly, students who received tutoring in Math reported being more comfortable in cross-cultural communication, although the effect was less pronounced (0.16/0.791 = 0.20 standard deviations). This is logical since there may be less interactions in the Math tutoring than in the English tutoring.

Table 6: Cross-Culture Communication

	(1)	(2)	(3)
	Index	Talking to someone from other country	Inverse: Worry when Talking to someone from other country
Math	0.16*	0.16**	0.13
	(0.08)	(0.08)	(0.11)
Math * School Closed	0.04	0.12	0.01
	(0.15)	(0.16)	(0.17)
School Closed	0.05	0.03	0.19
	(0.14)	(0.15)	(0.18)
English	0.41***	0.41***	0.40***

	(0.09)	(0.08)	(0.11)
Observations	821	822	820
R-squared	0.212	0.155	0.239
Mean Dep. Var.	3.922	3.943	3.900
SD	0.791	0.746	1.065

Note: Standard errors clustered at the student level in parentheses.. *** p<0.01, ** p<0.05, * p<0.1. Column 1 shows the estimation of equation 1 with the unweighted average of the cross-cultural communication index as the dependent variable. Columns 2 and 3 show the results of the same regression specification but with each individual component of the cross-cultural communication index as the dependent variable. The components of this index all range from 1 to 5, with 1 indicating the least amount of comfort and 5 indicating the highest level of comfort. They include: talking to someone from another country, and worrying about what to say when talking to someone from another country. Because column 3 asks a question where the ideal response is the lowest possible value, we reverse the response values (i.e. a response of 5 out of 5 for the question in column 3 now indicates that a student doesn't worry at all when talking to someone from another country). All regressions control for a student's age, gender, and current year of schooling, as well as the baseline survey response of the dependent variable and a dummy variable “Baseline Missing” equal to 1 if the baseline survey response is missing. If baseline survey response is missing, it is replaced by the value 0 and the dummy variable “Baseline Missing” takes the value 1.

In Appendix A, we also find a strong effect of both interventions (Math or English) on computer skills in Table 13. This is logical since for some students, this was the first time they were using a tablet with an internet connection.

4.6 Effects on Aspirations

We now turn to aspirations. Table 7 follows the same empirical specification, with the dependent variable in column 1 reflecting the standardized average of all questions in the Aspirations questionnaire and the remaining columns displaying the results for each individual component of the index.

Column 1 shows that the period of school closures is associated with a lower average Aspiration index score by 1.06 standard deviations. The coefficient of Math*SchoolClose d it is not significantly different from zero, indicating that the online tutoring program does not compensate for this loss in aspirations.

Table 7: Aspirations

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Index	Likely	Desired #	Spend on	Best	Desired	Motivated?	# of hours
		university	of kids	education?	job?	job?		to study
Math	0.15	0.14	-0.18*	-0.03	0.30**	0.33**	-0.02	0.12
	(0.10)	(0.09)	(0.10)	(0.10)	(0.15)	(0.15)	(0.15)	(0.11)

Math * School Closed	0.29	0.22	0.49*	-0.31	0.44	0.28	0.03	-0.02
	(0.25)	(0.21)	(0.27)	(0.22)	(0.39)	(0.37)	(0.15)	(0.15)
School Closed	-1.06***	-0.36*	-0.25	-0.93***	-0.80***	-0.60**	0.11	-0.61***
	(0.18)	(0.18)	(0.17)	(0.19)	(0.28)	(0.27)	(0.13)	(0.13)
English	-0.18**	-0.08	-0.08	-0.09	-0.16*	-0.15*	-0.04	-0.02
	(0.08)	(0.09)	(0.07)	(0.08)	(0.09)	(0.09)	(0.11)	(0.09)
Observations	822	822	777	823	744	722	820	818
R-squared	0.244	0.203	0.177	0.303	0.201	0.199	0.204	0.230
Mean Dep. Var	-0.362	-0.280	-0.171	-0.0455	-0.105	-0.107	-0.159	-0.321
SD	0.986	0.956	0.853	1.001	1.117	1.116	1.281	0.984

Note: Standard errors clustered at the student level in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Column (1) shows the estimation of equation 1 with the aspirations index as the dependent variable. In Column (2), the question is "How likely are you to go to university?" on a scale from 1 (definitely not go) to 5 (definitely will go. In Column (3), the dependent variable is the desired number of kids (inverted since we interpret a high response as having low aspirations; and standardized). In Column (4), the question is "If you were given 1000 Kenyan Shillings, how would you spend it?". Answers which related to school expenditures (e.g. bags, textbooks, uniforms, pens, pencils) were coded as 1 and other non-school related expenditures (e.g. toys, cell phone, radio, TV) were coded as 0. In Column (5), the question is: "What do you think is the best job in the world?". Answers which typically require higher education (e.g. doctor, nurse, engineer, lawyer) were tagged as 1 and other occupations (e.g. police man, soldier, football player) were tagged as 0. In Column (6), the question is: "Do you know what job you want to have in the future?". We re-code responses to this question in the same manner as above. In Column (7), the question is: "On a scale from 1 (not at all) to 10 (extremely motivated), how motivated are you to work hard?". In Column (8), the question is "How many hours per day would you be willing to spend on school work in order to go to university?". A high response to these previous two questions indicates high student aspirations. We standardize this variable. All variables in columns (2) to (8) are standardized, added together in an unweighted average, and re-standardized to make up the aspirations index of Column (1). All regressions control for a student's age, gender, and current year of schooling, as well as the baseline survey response of the dependent variable and a dummy variable "Baseline Missing" equal to 1 if the baseline survey response is missing. If baseline survey response is missing, it is replaced by the value 0 and the dummy variable "Baseline Missing" takes the value 1.

The rest of the table shows that this reduction in aspirations during the school closures comes from a decrease in aspirations to go to university (column 2), a reduction in the willingness to invest in one's education (column 4), a reduction in viewing high-skilled jobs as the best job or even a desirable job in the future (columns 5 and 6), and a reduction in the number of hours of work per day that they are willing to spend in order to go to university (column 8). School closures thus had a negative effect on aspirations.

The coefficient on the interaction of school closures and the treatment dummy isn't statistically different from 0. Thus, the tutoring program didn't salvage the lost aspirations experienced by the students due to the school closures.

In Appendix G, we present other indices such as motivation in Table 16, self-esteem in Table 17, and perceptions about Canada in Table 18 or Kenya in Table 19 and find very little effects of either interventions.

4.7 Discussion

The main result of the paper is that the online tutoring increases grades in Math when the schools are closed, but not when they are open. One explanation for this finding is decreasing returns to education.

In Appendix I, we use this fact to propose a methodology to estimate the learning loss, other than relying on difference-in-differences. We fit a model with decreasing returns to hours of math studied, using the exogenous variation provided by the randomized experiment implemented at two different points in time, after 0 hours studied (when the schools are closed) and 3 hours studied (when the schools are open). After estimating the model, we then use it to simulate school closures (i.e., going from 3 to 0 hours studied).

We find an estimate very close to the difference-in-difference estimator, and which does not rely on the parallel trends assumption. Instead, our estimator relies on a randomized experiment, implemented at two different points in time, such that we can evaluate the decreasing returns to hours of teaching in math in a production function of grades. The fact that these two methodologies yield relatively similar estimates support the claim that school closures causally created a large learning loss. This is important because most of the literature on quantifying the learning loss has been relying on a difference-in-differences estimate, which appears to be a valid estimator for the learning loss in our context.

5 Conclusion

School closures at the beginning of the COVID-19 pandemic had profound impacts on students' learning across the world. Governments around the world tried to put measures in place to address the learning loss. For example, in Kenya, the government introduced online distance learning initiatives through TV, radio, and internet uploads. These programs have been widely criticized by the literature for being inaccessible, especially in rural areas (Ochieng and Ngware, 2022; Malenya and Ohba, 2023; Mabeya, 2020). In this paper, we suggest the possibility of tutoring as an alternative. Tutoring can alleviate the concerns raised above: tutoring can be personalized at the right level, and it can reach even the rural underserved communities. Yet, no studies rigorously demonstrated the effects of online tutoring. Our paper is the first to do so. The policy implication of our paper is that tutoring can work as an alternative, especially when schools are closed.

Our study also adds to the literature about the effect of school closures on academic achievement. This has been the subject of an intense academic and policy debates, with estimates ranging from 0 to 0.7 SD, the higher estimates being found in remote rural areas (Singh, Romero and Muralidharan, 2022; Moscoviz and Evans, 2022; Patrinos, Vegas and Carter-Rau, 2022; Engzell, Frey and Verhagen, 2021; Maldonado and De Witte, 2020; Kuhfeld et al., 2020; Azevedo et al., 2020; Hevia et al., 2022). The fact that we collected data before and after the pandemic allows us to quantify the learning loss in our context. We compare the evolution in scores of the 2020 cohort to the 2019 one (in the control groups). We find a 0.8 SD reduction in education achievement test scores scores, on the high end of the estimates provided in the literature, which is consistent with the local context of a remote rural area of a developing country with few alternative online options available.

Our paper provides evidence for a new way to reduce this learning loss. Other strategies than online tutoring are currently being discussed to mitigate the learning loss of closing schools: SMS and 5-10-minute phone calls in Botswana (Angrist, Bergman and Matsheng, 2022), a similar program in Nepal (Radhakrishnan et al., 2021), 30-minute phone calls by teachers in Bangladesh (Beam, Mukherjee and Navarro-Sola, 2022), 30-minute phone tutoring sessions in Bangladesh (Hassan et al., 2022), teacher-student 15-minute mini-tutoring sessions in Kenya (Schueler and Rodriguez- Segura, 2021), and weekly phone tutorials from teachers in Sierra Leone (Crawfurd et al., 2022). The contribution of our paper is to study for the first time online video tutoring. We demonstrate that school closures led to significant learning loss (0.87 SD), 47% of which was compensated for by the tutoring program.

We also shed light on the heterogeneous effects of the tutoring program across time. While the program turned out to be a crucial part of students' education during lockdown, its impact on student grades in a normal time period was not statistically different from 0. This confirms the findings of a literature on tutoring that has found no effects when the schools are open (Nickow, Oreopoulos and Quan (2020) for non-professional volunteer tutors in after-school tutoring programs (the case in our paper), Romero, Chen and Magari (2021) for cross-age tutoring in Kenya, Ly, Maurin and Riegert (2020) in France, Kraft et al. (2022) for online tutoring in the US) but an effect when the schools are closed (Carlana and La Ferrara, 2021). This result is perhaps not very surprising ex-post; the marginal returns to an additional hour of tutoring are likely to be high when students aren't receiving any other education, but may be low if they are attending school full-time.

A limitation of our study is external validity since our sample is small and the intervention is implemented in rural Kenya. Reassuringly, our results are firmly within those of the existing literature in various different contexts in Italy (Carlana and La Ferrara, 2021), Botswana (Angrist, Bergman and Matsheng, 2022), Nepal (Radhakrishnan et al., 2021), Bangladesh (Beam, Mukherjee and Navarro-Sola, 2022; Hassan et al., 2022), Kenya (Schueler and Rodriguez-Segura, 2021), and Sierra Leone (Crawfurd et al., 2022).

Overall, we conclude that online tutoring can recover almost half of the cognitive losses, but none of the losses in aspirations. School closures had profound effects that must be fully understood and carefully estimated before closing schools.

References

Anderson, Michael L. 2008. “Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects.” Journal of the American Statistical Association, 103(484): 1481–1495.

Angrist, Noam, Peter Bergman, and Moitshepi Matsheng. 2022. “Experimental evidence on learning using low-tech when school is out.” Nature Human Behaviour, 6: 941–950.

Azevedo, Joao Pedro, Amer Hasan, Diana Goldemberg, Syedah Aroob Iqbal, and Koen Geven. 2020. Simulating the Potential Impacts of COVID-19 School Closures on Schooling and Learning Outcomes: A Set of Global Estimates.

Banerjee, Abhijit, Esther Duflo, Amy Finkelstein, Lawrence Katz, Benjamin Olken, and Anja Sautmann. 2020. “In Praise of Moderation: Suggestions for the Scope and Use of Pre-Analysis Plans for RCTs in Economics.” NBER Working Papers 26993, National Bureau of Economic Research, Inc.

Banerji, Rukmini, James Berry, and Marc Shotland. 2017. “The Impact of Maternal Liter- acy and Participation Programs: Evidence from a Randomized Evaluation in India.” American Economic Journal: Applied Economics, 9(4): 303–37.

BBC. 2021. “Coronavirus: Kenya reopens schools after nine months.”

BBC. 2022. “Uganda schools reopen after almost two years of Covid closure.”

Beam, Emily, Priya Mukherjee, and Laia Navarro-Sola. 2022. “Lowering Barriers to Remote Education: Experimental Impacts on Parental Responses and Learning.” IZA DP No. 15596.

Cameron, A Colin, Jonah B Gelbach, and Douglas L Miller. 2008. “Bootstrap-based improvements for inference with clustered errors.” The Review of Economics and Statistics, 90(3): 414–427.

Carlana, Michela, and Eliana La Ferrara. 2021. “Apart but Connected: Online Tutoring and Student Outcomes during the COVID-19 Pandemic.” Working Paper.

Chemin, Matthieu. 2018. “Informal Groups and Health Insurance Take-up Evidence from a Field Experiment.” World Development, 101: 54–72.

Crawfurd, Lee, David Evans, Susannah Hares, and Justin Sandefur. 2022. “Live Tutoring Calls Did Not Improve Learning during the COVID-19 Pandemic in Sierra Leone.” CGD Working Paper 591.

Education: From disruption to recovery. 2022.

Engzell, Per, Arun Frey, and Mark D. Verhagen. 2021. “Learning loss due to school closures during the COVID-19 pandemic.” Proceedings of the National Academy of Sciences, 118(17): e2022376118.

Glewwe, P., and K. Muralidharan. 2016. “Chapter 10 - Improving Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications.” In . Vol. 5 of Handbook of the Economics of Education, , ed. Eric A. Hanushek, Stephen Machin and Ludger Woessmann, 653–743. Elsevier.

Gore, Jennifer, Leanne Fray, Andrew Miller, Jess Harris, and Wendy Taggart. 2021. “The impact of COVID-19 on student learning in New South Wales primary schools: an empirical study.” The Australian Educational Researcher, 48.

Hanushek, Eric A., and Ludger Woessmann. 2020. “The economic impacts of learning losses.” OECD, , (225).

Hassan, Hashibul, Asad Islam, Abu Siddique, and Liang Choon Wang. 2022. “Telementoring and homeschooling during school closures: A randomized experiment in rural Bangladesh.” working paper.

Hevia, Felipe J., Samana Vergara-Lope, Anabel Veĺasquez-Duŕan, and David Calderón. 2022. “Estimation of the fundamental learning loss and learning poverty related to COVID-19 pandemic in Mexico.” International Journal of Educational Development, 88: 102515.

Hjort, Jonas, and Jonas Poulsen. 2019. “The Arrival of Fast Internet and Employment in Africa.” American Economic Review, 109(3): 1032–79.

Kraft, Matthew A., John A. List, Jeffrey A. Livingston, and Sally Sadoff. 2022. “Online Tutoring by College Volunteers: Experimental Evidence from a Pilot Program.” AEA Papers and Proceedings, 112: 614–618.

Kuhfeld, Megan, Beth Tarasawa, Angela Johnson, Erik Ruzek, and Karyn Lewis. 2020. “Learning during COVID-19: Initial findings on students’ reading and math achievement and growth.”

Ly, Son Thierry, Eric Maurin, and Arnaud Riegert. 2020. “A Pleasure That Hurts: The Ambiguous Effects of Elite Tutoring on Underprivileged High School Students.” Journal of Labor Economics, 38(2): 501–533.

Mabeya, Mary Theodorah. 2020. “Distance Learning During COVID-19 Crisis: Primary and Secondary School Parents Experiences in Kenya.” East African Journal of Education Studies, 2(1).

Maldonado, Joana, and Kristof De Witte. 2020. “The effect of school closures on standardised student test outcomes.” British Educational Research Journal.

Malenya, Francis Likoye, and Asayo Ohba. 2023. “Equity issues in the provision of online learning during the Covid-19 pandemic in Kenya Equity issues during online learning in Kenya” Journal of International Cooperation in Education.

McKenzie, David. 2012. “Beyond baseline and follow-up: The case for more T in experiments.” Journal of Development Economics, 99: 201–221.

Moscoviz, Laura, and David Evans. 2022. “Learning Loss and Student Dropouts during the COVID-19 Pandemic: A Review of the Evidence Two Years after Schools Shut Down.” CGD Working Paper 609.

Muris, Peter. 2001. “A Brief Questionnaire for Measuring Self-Efficacy in Youths.” Journal of Psychopathology and Behavioral Assessment, 23: 145–149.

Nickow, Andre, Philip Oreopoulos, and Vincent Quan. 2020. “The Impressive Effects of Tutoring on PreK-12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence.” Working Paper 27476 NBER.

Ochieng, Vollan Okoth, and Moses Waithanji Ngware. 2022. “Adoption of Education Technologies for Learning During COVID-19 Pandemic: The Experiences of Marginalized and Vulnerable Learner Populations in Kenya.” International Journal of Educational Reform, 1–24.

Oketch, Moses, and Maurice Mutisya. 2013. “Evolutional of educational outcomes in Kenya.” Paper commissioned for the EFA Global Monitoring Report 2013/4, Teaching and learning: Achieving quality for all.

Olanrewaju, Gideon Seun, Seun Bunmi Adebayo, Abiodun Yetunde Omotosho, and Charles Falajiki Olajidea. 2021. “2021; 2: 100092.Published online 2021 Nov 18. doi: 10.1016/j.ijedro.2021.100092PMCID: PMC8600108PMID: 35059671Left behind? The effects of digital gaps on e-learning in rural secondary schools and remote communities across Nigeria during the COVID19 pandemic.” International Journal of Educational Research Open.

Patrinos, Harry Anthony, Emiliana Vegas, and Rohan Carter-Rau. 2022. “An Analysis of COVID-19 Student Learning Loss.” World Bank Policy Research Working Paper 10033.

Pell, Tony, and Tina Jarvis. 2001. “Developing attitude to science scales for use with children of ages from five to eleven years.” International Journal of Science Education, 23(8): 847–862.

Radhakrishnan, Karthika, Shwetlena Sabarwal, Uttam Sharma, Claire Cullen, Colin Crossley, Thato Letsomo, and Noam Angrist. 2021. “Remote Learning: Evidence from Nepal during COVID-19.” World Bank Policy Brief.

Romero, Mauricio, Lisa Chen, and Noriko Magari. 2021. “Cross-Age Tutoring: Experimental Evidence from Kenya.” Economic Development and Cultural Change.

Rosenberg, Morris, Carmi Schooler, Carrie Schoenbach, and Florence Rosenberg. 1995. “Global Self-Esteem and Specific Self-Esteem: Different Concepts, Different Outcomes.” Ameri- can Sociological Review, 60(1): 141–156.

Schueler, Beth, and Daniel Rodriguez-Segura. 2021. “A Cautionary Tale of Tutoring Hard- to-Reach Students in Kenya.” EdWorkingPaper: 21-432.

Schult, Johannes, Nicole Mahler, Benjamin Fauth, and Marlit A Lindner. 2021. “Did Students Learn Less During the COVID-19 Pandemic? Reading and Mathematics Competencies Before and After the First Pandemic Wave.”

Singh, Abhijeet, Mauricio Romero, and Karthik Muralidharan. 2022. “COVID-19 Learning Loss and Recovery: Panel Data Evidence from India.” National Bureau of Economic Research Working Paper 30552.

Webb, Matthew. 2014. “Reworking Wild Bootstrap Based Inference for Clustered Errors.” Eco- nomics Department, Queen’s University.

Woessmann, Ludger, Vera Freundl, Elisabeth Grewenig, Philipp Lergetporer, Katharina Werner, and Larissa Zierow. 2020. “Education in the corona crisis: How did the schoolchildren spend the time the schools were closed and which educational measures do the Germans advocate?” ifo Institute.

Young, Alwyn. 2018. “Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results.” The Quarterly Journal of Economics, 134(2): 557–598.

APPENDIX

A Appendix A: English Proficiency Measure

To measure oral proficiency in English (which is not assessed in the exams), we use a test constructed to be mapped into international language standards. Native English speakers were hired and paid by an external organization (called the "McGill Arts Internship Office") to physically travel to the site of the research project in Kenya. These interns were not tutors themselves and were blind to the experiment in the sense that they were never shown the randomized list of who was in the treatment or control group.[18]

These interns were all native English speakers and were thus able to gauge oral proficiency in English. They ask seven questions to start and facilitate a conversation. The first questions are easy with concrete subjects and a familiar vocabulary (i.e., Do you prefer rice or ugali?), while the last questions are harder with more abstract subjects (i.e., Can you describe for me the meaning of the word kindness?).[19]

Considering the range of questions, the test is designed to be informative over a wide range of student achievement.

These questions are different from those suggested for the ice-breaking activities of the tutors. The tutors were never informed about the content of this oral proficiency test such that it would not have been possible for them to teach to the test. In any case, tutors had no incentives to teach to the test, they were entirely volunteering their time with no rewards being given for certain results.

These questions were carefully chosen after extensive piloting to deal with issues of time and shyness. The reasoning behind them was that asking students more direct questions elicited more direct answers. In a previous version of the test, we showed cartoons and asked students to describe them, followed by a storytelling/listening activity. The open-endedness of the photo-based questions struck students silent – even those that spoke English well. After that, it was hard to refocus the conversation, and the interview became awkward. This obviously only made students clam up more. We discovered it was easier to ask a question, see what happens, and continue. The pictures were overwhelming. It was also hard to find cartoons that both

[18] The main occupation of these interns was to develop their own independent research project (different from this project), collect their own data, analyze it and produce a working paper for academic credits on their return to the university. To get experience collecting data, they collected these oral proficiency tests. These interns were not paid by the experimenter.

[19] The full list is: 1. What is your name? How old are you? Do you have any brothers or sisters? Can you tell me about them? (Finding out basic personal information, warm-up questions.) 2. Do you prefer rice or ugali? Why is that? (Warm-up, concrete subject, familiar vocabulary, likes/dislikes.) 3. Do you have a musician or television program? Can you tell me about it/them? Why do you like it/them? (Concrete subject, likes/dislikes, and opportunity to demonstrate range of vocabulary and fluency.) 4. Can you name a sport you would like to play one day? A food you would like to try? A place you would like to visit? (Concrete subject, less familiar vocabulary, uses future tense.) 5. Can you describe for me the meaning of the word kindness? (Abstract subjects.) 6. Can you think of an occasion where you were very happy? Can you tell me about it? (Abstract subjects, past tense.) 7. I want you to try to think of a question to ask me. It can be about anything! (Ability to ask questions.)

suited the context and had enough activity going on. The storytelling/listening activity made the test too long. The students have limited attention spans, and once they lost interest or sat in silence for too long, it was hard to get them back on track. The test used in this paper with a few direct questions deals with these issues of time and shyness. The beginning conversational questions get students comfortable and give them time to warm up. Having pictures to look at and things to listen to made it feel like more of a "test", whereas the few questions is more of a casual "chit chat." In this way, the native English speakers were able to elicit responses from students and gauge their level of oral proficiency.

The native English speakers then grade each student on four different dimensions: understanding a native speaker, conversation, vocabulary range, and spoken fluency. They use a "rubric", i.e., in education terminology, a scoring guide used to evaluate the quality of students' constructed responses, established by the "Common European Framework of Reference for Languages (CEFR)", put together by the Council of Europe as a way of standardizing the levels of language exams in different regions. The CEFR scoring rubrics are important since they are widely used internationally and all important exams are mapped to them.[20]

The rubrics for the Oral Proficiency Test are:

- Understanding a native speaker

Can understand any native speaker, even on abstract and complex topics of a specialist nature beyond his/her own field, given an opportunity to adjust to a non-standard accent or dialect.
Can understand in detail speech on abstract and complex topics of a specialist nature beyond his/her own field, though he/she may need to confirm occasional details, especially if the accent is unfamiliar.
Can understand in detail what is said to him/her in the standard spoken language even in a noisy environment.
Can follow clearly articulated speech directed at him/her in everyday conversation, though will sometimes have to ask for repetition of particular words and phrases.
Can understand enough to manage simple, routine exchanges without undue effort. Can generally understand clear, standard speech on familiar matters directed at him/her, provided he/she can ask for repetition or reformulation from time to time.
Can understand everyday expressions aimed at the satisfaction of simple needs of a concrete type, delivered directly to him/her in clear, slow and repeated speech by a sympathetic speaker. Can understand questions and instructions addressed carefully and slowly to him/her and follow short, simple directions.

- Conversation

Can converse comfortably and appropriately, unhampered by any linguistic limitations in conducting a full social and personal life.

[20] See for more details: https://www.coe.int/en/web/common-european-framework-reference-languages/home

Can use language flexibly and effectively for social purposes, including emotional, allusive and joking usage.
Can engage in extended conversation on most general topics in a clearly participatory fashion, even in a noisy environment. Can sustain relationships with native speakers without unintentionally amusing or irritating them or requiring them to behave other than they would with a native speaker. Can convey degrees of emotion and highlight the personal significance of events and experiences.
Can enter unprepared into conversations on familiar topics. Can follow clearly articulated speech directed at him/her in everyday conversation, though will sometimes have to ask for repetition of particular words and phrases. Can maintain a conversation or discussion but may sometimes be difficult to follow when trying to say exactly what he/she would like to. Can express and respond to feelings such as surprise, happiness, sadness, interest and indifference.
Can establish social contact: greetings and farewells; introductions; giving thanks. Can generally understand clear, standard speech on familiar matters directed at him/her, provided he/she can ask for repetition or reformulation from time to time. Can participate in short conversations in routine contexts on topics of interest. Can express how he/she feels in simple terms, and express thanks. Can handle very short social exchanges but is rarely able to understand enough to keep conversation going of his/her own accord, though he/she can be made to understand if the speaker will take the trouble. Can use simple everyday polite forms of greeting and address. Can make and respond to invitations, suggestions and apologies. Can say what he/she likes and dislikes.
Can make an introduction and use basic greeting and leave-taking expressions. Can ask how people are and react to news. Can understand everyday expressions aimed at the satisfaction of simple needs of a concrete type, delivered directly to him/her in clear, slow and repeated speech by a sympathetic speaker.

- Vocabulary range

Has a good command of a very broad lexical repertoire including idiomatic expressions and colloquialisms; shows awareness of connotative levels of meaning.
Has a good command of a broad lexical repertoire allowing gaps to be readily overcome with circumlocutions; little obvious searching for expressions or avoidance strategies. Good command of idiomatic expressions and colloquialisms.
Has a good range of vocabulary for matters connected to his/her field and most general topics. Can vary formulation to avoid frequent repetition, but lexical gaps can still cause hesitation and circumlocution.
Has a sufficient vocabulary to express him/herself with some circumlocutions on most topics pertinent to his/her everyday life such as family, hobbies and interests, work, travel, and current events.
Has sufficient vocabulary to conduct routine, everyday transactions involving familiar situations and topics. Has a sufficient vocabulary for the expression of basic communicative needs. Has a sufficient vocabulary for coping with simple survival needs.
Has a basic vocabulary repertoire of isolated words and phrases related to particular concrete situations.

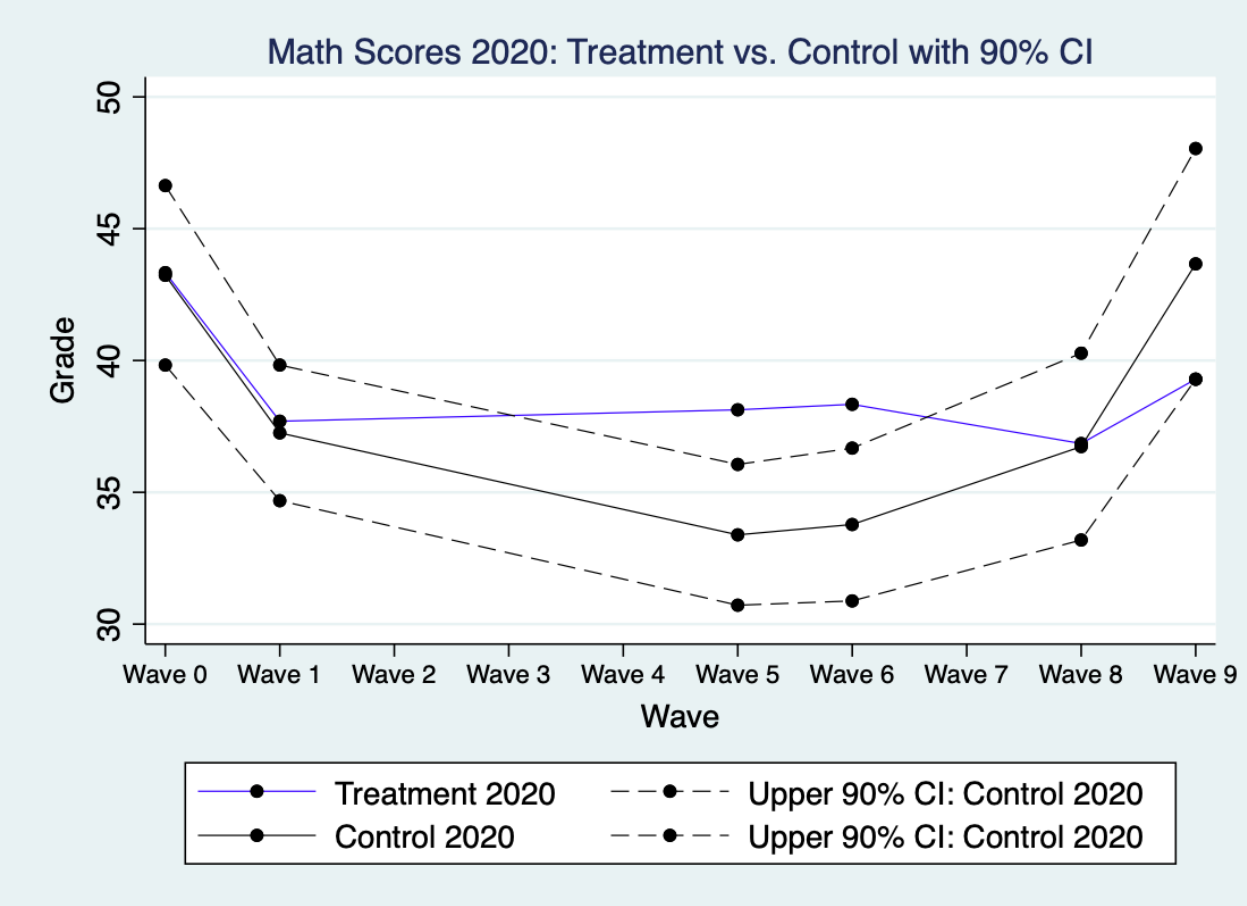
- Spoken fluency

Can express him/herself at length with a natural, effortless, unhesitating flow. Pauses only to reflect on precisely the right words to express his/her thoughts or to find an appropriate example or explanation.
Can express him/herself fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language.
Can communicate spontaneously, often showing remarkable fluency and ease of expression in even longer complex stretches of speech. Can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he/she searches for patterns and expressions, there are few noticeably long pauses. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without imposing strain on either party.
Can express him/herself with relative ease. Despite some problems with formulation resulting in pauses and 'cul-de-sacs', he/she is able to keep going effectively without help. Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production.
Can make him/herself understood in short contributions, even though pauses, false starts and reformulation are very evident. Can construct phrases on familiar topics with sufficient ease to handle short exchanges, despite very noticeable hesitation and false starts.
Can manage very short, isolated, mainly pre-packaged utterances, with much pausing to search for expressions, to articulate less familiar words, and to repair communication.

B Appendix B: Confidence Intervals

in Figure 1, we display the confidence intervals for the control group in the year 2020. The average maths grade for the treatment group is above the upper bound of the 90% confidence interval in Waves 5 and 6, the waves directly following the reopening of schools.
The average maths grade for the treatment group remains within the confidence interval in Wave 9, indicating that the large spike observed for the control group in wave 9 is not significantly different from the treatment group.
We conclude that the only significant difference is in waves 5 and 6, not in other waves.

Figure 2: Math Grades: Treatment vs Control in times of COVID-19



Note: The figure shows trends across the 9 waves within the school year of 2020, split by treatment and control groups. The dashed lines indicate the 90% confidence interval for the control group. Schools were closed for waves 2 through 5 in 2020, but online tutoring continued.

C Appendix C: Control Variables

Table 8 includes a student's baseline value for a specific index of survey questions. Specifically, we control for baseline English oral comprehension in Column (1), computer proficiency in Column (2), cross-culture communication in Column (3), motivation in Column (4), self-esteem in Column (5), aspiration in Column (6), liking school in Column (7), liking courses in Column (8), thoughts on Canada in Column (9), and thoughts on Kenya in Column (10) . We find similar results in all columns.

Table 8: Math grades: 2016-2018 vs 2020, Index Controls

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Dependent Variable: Math grade									
Math	-0.57	-0.65	-1.01	-0.43	-0.65	-0.59	-0.74	-0.68	-0.68	-0.66
	(1.25)	(1.20)	(1.21)	(1.18)	(1.20)	(1.23)	(1.24)	(1.23)	(1.20)	(1.20)

Math * School Closed	6.09**	5.50**	6.05**	5.53**	5.65**	5.60**	5.67**	5.67**	5.57**	5.65**
	(2.83)	(2.72)	(2.67)	(2.68)	(2.70)	(2.70)	(2.70)	(2.70)	(2.69)	(2.69)
School Closed	-8.31***	-12.88***	-10.54***	-13.43***	-12.10***	-12.47***	-12.05***	-12.07***	-11.67***	-12.19***
	(2.30)	(2.13)	(2.20)	(2.22)	(2.20)	(2.16)	(2.17)	(2.19)	(2.31)	(2.25)
Wave 5 * 2016-2018	-0.23	-0.22	-0.24	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22	-0.22
	(1.12)	(1.13)	(1.13)	(1.12)	(1.13)	(1.13)	(1.13)	(1.13)	(1.13)	(1.12)
English	-1.43	-1.11	-1.12	-0.79	-1.03	-0.95	-1.06	-1.06	-1.06	-1.05
	(1.35)	(1.40)	(1.37)	(1.39)	(1.42)	(1.39)	(1.40)	(1.41)	(1.41)	(1.40)
Control	Oral	Computer	X-Culture	Motiv.	Self-	Aspiration	Like	Like	Canada	Kenya
	Comp.	Prof.	Comm.		Esteem		School	Courses		
Observations	2,170	2,170	2,170	2,170	2,170	2,170	2,170	2,170	2,170	2,170
R-squared	0.368	0.357	0.363	0.364	0.356	0.357	0.356	0.356	0.356	0.356
Mean Dep. Var	40.82	40.82	40.82	40.82	40.82	40.82	40.82	40.82	40.82	40.82
SD	13.77	13.77	13.77	13.77	13.77	13.77	13.77	13.77	13.77	13.77

Note: Standard errors clustered at the student level in parentheses.. *** p<0.01, ** p<0.05, * p<0.1. For each column, the dependent variable is a student's math grade in a given wave. “Math Treatment” is a dummy equal to 1 if a student is in the Math treatment group. “School Closed” is a dichotomous variable equal to 1 when schools are closed, i.e., wave 5 of 2020. “MathTreatment * School Closed” is the interaction between the two variables.. “Wave 5 * 2016-2018” is a dummy equal to 1 if the period is wave 5 in any of the years 2016 - 2018. “English” is a dummy variable equal to 1 student i was in the treatment group for the years 2016 to 2018. All regressions include a full set of interactions between the 9 waves and the 3 time periods (2016-2018 for the English treatment, 2019 for the Math treatment, and 2020 for the Math treatment combined with school closure). All regressions control for baseline math grade and a dummy variable “Baseline Missing” equal to 1 if the baseline math grade is missing. If baseline math grade is missing, it is replaced by the value 0 and the dummy variable “Baseline Missing” takes the value 1. Each column includes a student's baseline value for a specific index of survey questions. For columns 1-10, these indices include respectively: oral comprehension, computer proficiency, cross-culture communication, motivation, self-esteem, future aspirations, liking school in general, liking courses, thoughts about living in Canada, and thoughts about living in Kenya.

Table 9 includes a student's baseline value for the 51 components of the 10 sections of the survey.

Table 9: All Components

	Maths
Math	0.23
	(1.14)
Math * School Closed	4.85*
	(2.50)
School Closed	-28.26***
	(4.81)
English	-1.31
	(1.21)
Control Variable	All Components
Observations	2,113
R-squared	0.483
Mean Dep. Var.	40.82
SD	13.77

Note: Standard errors clustered at the student level in parentheses. *** p<0.01, ** p<0.05, * p<0.1. This column includes the baseline value of the 51 components of the 10 sections of the survey.

D Appendix D: Disaggregating wave*year fixed effects

Our specification in Table 3 groups together the years 2016-2018 within the wave*year fixed effects. In Table 10 below, we use the same specification from Table 3- the only difference is that we relax the above assumption by disaggregating the wave*year fixed effects. This has essentially no impact on the results from Table 3. Column (1) of Table 10 shows that the variable Mat h it is still not statistically different from 0, indicating that the educational program did not have any significant effect on math grades. When the schools are closed, the coefficient of Math*SchoolClose d it is statistically significant at the 5% level and now has a coefficient of 5.87, indicating that being in the treatment group was associated with a math score 5.87 points higher compared to the control group.

Table 10: Math grades: Disaggregated Wave*Year Fixed Effects for the 2016-2018 period

	(1)	(2)	(3)	(4)
	Dependent Variable: Math grade			

Math	-0.84	-0.82	-0.59	-0.58
	(1.28)	(1.21)	(1.20)	(1.16)
Math * School Closed	5.87**	6.39**	6.54**	6.12**
	(2.76)	(2.72)	(2.79)	(2.77)
School Closed	-10.67***	-10.90***	-11.64***	-11.08**
	(2.11)	(2.23)	(2.44)	(4.35)
Wave*Year fixed effects	YES	YES	YES	YES
Controls:				
Baseline grade	NO	YES	YES	YES
Age, Gender, School Year	NO	NO	YES	YES
Baseline Survey	NO	NO	NO	YES
Observations	2,170	2,170	2,170	2,170
R-squared	0.537	0.560	0.564	0.572
Mean dep var	40.82	40.82	40.82	40.82
SD dep var	13.77	13.77	13.77	13.77

Note: Standard errors clustered at the student level in parentheses. *** p<0.01, ** p<0.05, * p<0.1. For each column, the dependent variable is a student's math grade in a given wave. 'Math' is a dummy equal to 1 if a student was in the Math treatment group for the years 2019-2020. 'School Closed' is a dichotomous variable equal to 1 when schools are closed, i.e., wave 5 of 2020. 'MathTreatment * School Closed' is the interaction between the two variables. All regressions include a full set of interactions between the 9 waves and the 5 years. All regressions control for baseline math grade and a dummy variable 'Baseline Missing' that is equal to 1 if the baseline math grade is missing. If baseline math grade is missing, it is replaced by the value 0 and the dummy variable 'Baseline Missing' takes the value 1. Column 1 shows the estimation of Equation 1 without any additional controls. Column 2 augments this specification by including the baseline grade on all other topics than Maths, and a dummy variable 'Baseline Missing Total grade' equal to 1 if the baseline total grade is missing. Column 3 adds to column 2 by controlling for a student's age, gender, and the year of schooling they are currently completing. Finally, column 4 includes in the list of controls the baseline averages of various indices from the survey data. These indices include: oral comprehension, computer proficiency, cross-culture communication, motivation, self-esteem, future aspirations, liking school, liking classes, thoughts about Canada, and thoughts about Kenya.

E Appendix E : Estimation with 2019-2020

In Table 11 below, we restrict the sample to the years 2019 and 2020 alone when the tutoring was in Maths. The results are very similar to the main results of the paper.

Table 11: Math Grades: 2019 and 2020

	(1)	(2)	(3)	(4)
	Dependent Variable: Math Grade			
Math	-0.68	-0.53	-0.09	-0.44
	(1.20)	(1.09)	(1.13)	(1.04)
Math * School Closed	5.66**	6.10**	6.30**	6.31**
	(2.71)	(2.64)	(2.74)	(2.70)
School Closed	-7.31***	-7.15***	-6.51***	-8.14**
	(1.87)	(1.87)	(1.94)	(3.50)
Wave*Year fixed effects	YES	YES	YES	YES
Controls:				
Baseline Grade	NO	YES	YES	YES
Age, Gender, School Year	NO	NO	YES	YES
Baseline Survey	NO	NO	NO	YES
Observations	867	867	867	867
R-squared	0.439	0.481	0.488	0.503
Mean Dep. Var	37.51	37.51	37.51	37.51
SD	12.68	12.68	12.68	12.68

Note: Standard errors clustered at the student level in parentheses. *** p<0.01, ** p<0.05, * p<0.1. For each column, the dependent variable is a student's math grade in a given wave. 'Math' is a dummy equal to 1 if a student was in the Math treatment group for the years 2019-2020. 'School Closed' is a dichotomous variable equal to 1 when schools are closed, i.e., wave 5 of 2020. 'MathTreatment * School Closed' is the interaction between the two variables. All regressions include a full set of interactions between the 9 waves and the 2 time periods (2019 and 2020). All regressions control for baseline math grade and a dummy variable 'Baseline Missing' that is equal to 1 if the baseline math grade is missing. If baseline math grade is missing, it is replaced by the value 0 and the dummy variable 'Baseline Missing' takes the value 1. Column 1 shows the estimation of Equation 1 without any additional controls. Column 2 augments this specification by including the baseline grade on all other topics than Maths, and a dummy variable 'Baseline Missing Total Grade' equal to 1 if the baseline total grade is missing. Column 3 adds to column 2 by controlling for a student's age, gender, and the year of schooling they are currently completing. Finally, column 4 includes in the list of controls the baseline averages of various indices from the survey data. These indices include: oral comprehension, computer proficiency, cross-culture communication, motivation, self-esteem, future aspirations, liking school, liking classes, thoughts about Canada, and thoughts about Kenya.

F Appendix F: Attrition Test

We create a dummy variable that represents a student's attrition status for a given wave-year. If the student is missing the math grade, the variable is set to 1. We first show that the group of students receiving the Math intervention is not associated with a statistically different attrition status in Column (1). We then add in the "School Closed" dummy and its interaction with the Math intervention in Column (2) , as well as baseline total grade, age, gender, school year, and baseline survey responses. Reassuringly, neither the School Closed period nor the interaction term are associated with higher or lower attrition.

Table 12: Attrition Test

	(1)	(2)
	Dependent Variable: Attrition	
Math	-0.00	-0.01
	(0.01)	(0.01)
Math * School Closed		-0.06
		(0.04)
School Closed		0.03
		(0.04)
Wave 5 * 2016-2018	-0.01	-0.01
	(0.02)	(0.02)
English	-0.03**	-0.02**
	(0.01)	(0.01)
Wave*Year fixed effects	YES	YES
Controls:		
Baseline grade	NO	YES
Age, Gender, School Year	NO	YES
Baseline Survey	NO	YES
Observations	2,212	2,212
R-squared	0.040	0.105
Mean dep var	0.019	0.019
SD dep var	0.138	0.138

Note: Standard errors clustered at the student level in parentheses. *** p<0.01, ** p<0.05, * p<0.1. For each column, the dependent variable is a student's attrition status in a given wave. 'Math' is a dummy equal to 1 if a

student was in the Math treatment group for the years 2019-2020. 'School Closed' is a dichotomous variable equal to 1 when schools are closed, i.e., wave 5 of 2020. 'MathTreatment * School Closed' is the interaction between the two variables. 'Wave 5 * 2016-2018' is a dummy equal to 1 if the period is wave 5 in any of the years 2016 - 2018. 'English' is a dummy variable equal to 1 if a student was in the treatment group for the years 2016 - 2018. All regressions include a full set of interactions between the 9 waves and the 3 time periods (2016-2018 for the English treatment, 2019 for the Math treatment, and 2020 for the Math treatment combined with school closure).

G Appendix G: Other Outcomes

G.0.1 Computer Proficiency

In the Computer Proficiency index, we ask students a series of questions about their comfort with using computers and technology. These include: “How comfortable do you feel using a computer, including the internet?”; “How comfortable do you feel using the internet on a computer?”; “How comfortable do you feel using the internet on a cell phone?”; “How comfortable do you feel sending an email?”; “How comfortable do you feel talking on Skype?” All questions range from 1 to 5, with 5 being the highest comfort level.

Table 13: Computer Proficiency

	(1)	(2)	(3)	(4)	(5)	(6)
	Index	Using	Using Internet	Using internet	Using	Using
		computer	on computer	on phone	email	video call
Math	0.80***	0.84***	0.85***	0.42***	0.43***	1.75***
	(0.11)	(0.15)	(0.15)	(0.10)	(0.11)	(0.15)
Math * School Closed	0.31	0.31	0.30	0.23	0.54**	-0.27
	(0.20)	(0.27)	(0.26)	(0.20)	(0.23)	(0.21)
School Closed	0.87***	0.66**	1.08***	1.46***	0.84***	1.03***
	(0.25)	(0.27)	(0.37)	(0.28)	(0.29)	(0.34)
English	0.87***	0.96***	0.43***	0.22**	0.02	2.69***
	(0.06)	(0.11)	(0.09)	(0.10)	(0.04)	(0.12)
Observations	819	811	808	811	803	604
R-squared	0.545	0.285	0.443	0.586	0.440	0.648
Mean Dep. Var	2.228	2.476	1.996	2.308	1.445	3.280
SD	0.976	1.199	1.206	1.313	0.843	1.392

Note: Standard errors clustered at the student level in parentheses.. *** p<0.01, ** p<0.05, * p<0.1. Each component goes from 1 to 5, with 5 being the most comfort. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Column 1 shows the estimation of equation 1 with the unweighted average of the computer

proficiency index as the dependent variable. Columns 2 through 5 show the results of the same regression specification but with each individual component of the computer proficiency index as the dependent variable. The components of this index all range from 1 to 5, with 1 indicating the least amount of comfort and 5 indicating the highest level of comfort. They include: using a computer, using internet on a computer, using internet on a phone, using email, and using video call. All regressions control for a student's age, gender, and current year of schooling, as well as the baseline survey response of the dependent variable and a dummy variable "Baseline Missing" equal to 1 if the baseline survey response is missing. If baseline survey response is missing, it is replaced by the value 0 and the dummy variable "Baseline Missing" takes the value 1.
Table 13 shows the estimation of equation 1 with the unweighted average of the computer proficiency index as the dependent variable, followed by each individual question. Both interventions (Math and English) significantly increase the score, by a large amount (0.8 and 0.87 out of 5). Notice that the coefficient of SchoolClosed is positive. One potential explanation for this seemingly counterintuitive result is that the positive coefficient is a reflection of the general trend that students become more proficient with technology over time. Indeed, other (omitted) wave * year fixed effects also show positive coefficients. Since the omitted wave is "Wave 1 * 2016-2018" (i.e., the very first wave in the sample), all other wave * year fixed effects capture student-invariant time trends later in time.

G.0.2 Liking School

Table 14 shows the components of the Liking School index. We use modified questions from Pell and Jarvis (2001) and asked students how they felt about doing or learning certain subjects in school. In each column, the question asked to students is: "How do you feel about learning this subject/doing this activity related to school?", where student answers vary from 1 (don't like at all) to 5 (really like). They include: english composition, learning english, mathematics, christian religion, social studies, science, insha, and swahili.
School closures improve the scores for almost each field. In other words, students declare that they like school when schools are closed.
The interaction of the math tutoring program and school closures has no impact on liking school: the tutoring program does not have a differential effect on the Liking School index over and above the school closures.

Table 14: Liking School

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	English Comp	Learning English	Math	Christian Religion	Social Studies	Science	Insha	Swahili
Math	-0.05	-0.01	-0.03	-0.12*	-0.03	0.01	0.05	0.12
	(0.09)	(0.08)	(0.10)	(0.06)	(0.11)	(0.11)	(0.09)	(0.08)
Math * School Closed	0.02	-0.07	0.12	0.40***	-0.09	-0.20	-0.28	-0.09
	(0.23)	(0.21)	(0.28)	(0.15)	(0.23)	(0.24)	(0.21)	(0.19)
School Closed	0.32*	0.27	0.11	0.48***	0.18	0.40***	0.72***	0.58***
	(0.17)	(0.17)	(0.21)	(0.12)	(0.16)	(0.15)	(0.14)	(0.14)
English	-0.01	0.18**	-0.04	-0.07	0.02	-0.03	-0.09	0.06
	(0.08)	(0.09)	(0.09)	(0.08)	(0.09)	(0.07)	(0.08)	(0.07)

Observations	820	821	822	820	822	822	822	821
R-squared	0.200	0.140	0.094	0.193	0.116	0.129	0.164	0.192
Mean Dep. Var	3.830	4.041	3.912	4.045	3.691	4.030	3.811	4.004
SD	0.755	0.762	0.829	0.682	0.840	0.737	0.782	0.748

Note: Standard errors clustered at the student level in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Column 1 shows the estimation of equation 1 with the unweighted average of the self-esteem index as the dependent variable. Columns 2-9 go from 1-5. In each column, the question asked to students is: "How do you feel about learning this subject/doing this activity related to school?", where student answers vary from 1 (don't like at all) to 5 (really like). They include: english composition, learning english, mathematics, christian religion, social studies, science, insha, and swahili. All regressions control for a student's age, gender, and current year of schooling, as well as the baseline survey response of the dependent variable and a dummy variable "Baseline Missing" equal to 1 if the baseline survey response is missing. If baseline survey response is missing, it is replaced by the value 0 and the dummy variable "Baseline Missing" takes the value 1.

Table 15 shows the results of these regressions for the Liking School index. We use an unweighted average index of all subjects in Column 1: school closures improve the liking school index.

Table 15: Liking School

	(1)	(2)	(3)	(4)
	Index	Working alone	Working with others	Coming to school
Math	-0.05	-0.14*	-0.29***	-0.12**
	(0.04)	(0.09)	(0.10)	(0.05)
Math * School Closed	0.02	0.26**	0.39**	0.00
	(0.11)	(0.11)	(0.17)	(0.11)
School Closed	0.23***	-1.68***	0.48***	0.76***
	(0.07)	(0.13)	(0.13)	(0.07)
English	0.01	0.04	0.11	-0.01
	(0.03)	(0.09)	(0.07)	(0.04)
Observations	821	821	822	820
R-squared	0.233	0.515	0.200	0.315
Mean Dep. Var	3.897	3.028	4.133	4.343
SD	0.362	1.136	0.854	0.538

Note: Standard errors clustered at the student level in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Column 1 shows the estimation of equation 1 with the unweighted average of the liking school index as the dependent variable. For each subject, the question asked to students is: "How do you feel about learning this subject/doing this activity related to school?", where student answers vary from 1 (don't like at all) to 5 (really like). They include: English composition, learning English, mathematics, christian religion, social studies, science, insha, and swahili. In Column 2, the dependent variable is the answer to the question: "How do you feel about working by yourself at school?". In Column 3, the question is "How do you feel about working with others at school?", and in column 4 "How do you feel about coming to school?". Student answers vary from 1 (don't like at all) to 5 (really like). All regressions control for a student's age, gender, and current year of schooling, as well as the baseline survey response of the dependent variable and a dummy variable "Baseline Missing" equal to 1 if the baseline survey response is missing. If baseline survey response is missing, it is replaced by the value 0 and the dummy variable "Baseline Missing" takes the value 1.

In Column 2, we ask: "How do you feel about working by yourself at school?". When the schools are closed, students report liking less working alone. Column 3 shows on the other hand that students prefer working with others. Column 4 shows that student feel better about coming to school (the exact question is: "How do you feel about coming to school?", once again when the schools are closed.
Intuitively, the two results on loss of aspirations and liking school go hand-in-hand: students like school, but schools are closed - this hurts students' aspirations because they know it will be harder to go to university and get high-skilled jobs.

G.0.3 Motivation

We use questions from Muris (2001) for the section on academic motivations, where each component in the Motivation index asks the student how well he or she can do on a certain task related to motivation (i.e. column 2 asks "How well can you get help when stuck on homework?"). The components all range from 1 to 5, with 1 indicating a very low ability to complete the task and 5 indicating a very strong ability. They include: getting help when stuck on homework, studying when there are other interesting things, doing revision before an exam, succeeding in finishing all your homework everyday, paying attention during every class, succeeding in passing courses, parents being satisfied with school performance, and easily passing a test.
Table 16 displays the results for the academic motivations module in the survey where students were asked this series of questions related to their school habits. Despite documenting an overall positive effect of school closures on academic motivation, we find varying results across the components of the index: some components are positively affected (columns 5, 7, 8, 9) while others are negatively affected (columns 2 and 3). It is thus difficult to conclude that motivation is affected in a single direction.

Table 16: Motivation

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Index	Get help	Study	Revision	Finish	Pay	Passing	School	Pass
					homework	attention	courses	performance	test

Math	0.05	-0.01	0.06	0.05	0.09	0.03	0.03	0.09	-0.01
	(0.06)	(0.10)	(0.12)	(0.12)	(0.12)	(0.09)	(0.09)	(0.11)	(0.10)
Math * School Closed	0.11	0.11	-0.15	0.24	0.02	0.09	0.03	0.25	0.25
	(0.11)	(0.20)	(0.24)	(0.20)	(0.16)	(0.15)	(0.17)	(0.21)	(0.18)
School Closed	0.29***	-0.98***	-0.37**	-0.24	0.43***	0.16	0.96***	0.78***	1.82***
	(0.11)	(0.20)	(0.17)	(0.19)	(0.15)	(0.13)	(0.15)	(0.17)	(0.18)
English	0.01	0.04	-0.08	0.05	0.05	-0.06	-0.07	0.04	0.04
	(0.05)	(0.14)	(0.10)	(0.09)	(0.09)	(0.08)	(0.07)	(0.09)	(0.08)
Observations	821	819	819	820	822	822	822	821	821
R-squared	0.345	0.316	0.135	0.096	0.119	0.107	0.483	0.420	0.652
Mean Dep. Var.	3.258	2.938	2.573	3.840	3.791	3.940	3.294	2.903	2.781
SD	0.528	1.155	0.960	0.795	0.779	0.712	0.882	1.084	1.254

Note: Standard errors clustered at the student level in parentheses.. *** p<0.01, ** p<0.05, * p<0.1. Each component goes from 1 to 5, with 5 being the most, and each question asks "How well can you...." Column 1 shows the estimation of equation 1 with the unweighted average of the motivation index as the dependent variable. Columns 2 and 3 show the results of the same regression specification but with each individual component of the motivation index as the dependent variable. Each component in this index asks the student how well he or she can do on a certain task related to motivation (i.e. column 2 asks "How well can you get help when stuck on homework?"). The components all range from 1 to 5, with 1 indicating a very low ability to complete the task and 5 indicating a very strong ability, and include: getting help when stuck on homework, studying when there are other interesting things, doing revision before an exam, succeeding in finishing all your homework every day, paying attention during every class, succeeding in passing courses, parents being satisfied with your school performance, and easily passing a test. All regressions control for a student's age, gender, and current year of schooling, as well as the baseline survey response of the dependent variable and a dummy variable "Baseline Missing" equal to 1 if the baseline survey response is missing. If baseline survey response is missing, it is replaced by the value 0 and the dummy variable "Baseline Missing" takes the value 1.

Additionally, the interpretation of some of the questions is challenging during the school closure period. For example, students report struggling more during school closures with getting help when stuck on homework while also reporting that they succeed more in passing their courses during the school closures. Yet, there were no homework assignments or school tests during the period in which schools were closed.

G.0.4 Self-Esteem

Next, we use questions from Rosenberg et al. (1995) related to student self-esteem. The statements are: "I am satisfied with myself", "I think I am no good at all", "I feel that I have a number of good qualities", "I am able to do things as well as most others", "I feel I do not have much to be proud of", "I certainly feel useless at times", "I feel that I am a person of worth", "I wish I could have more respect for myself", "I sometimes feel that I'm a failure", and "I take a positive attitude toward myself". Answers range from 1 to 4, with 4 being strongly agree and 1

strongly disagree. Because columns 3, 6, 7, 9, and 10 ask questions where the ideal response is the lowest possible value, we reverse the response values (i.e. a response of 4 out of 4 for the question in column 6 now indicates that a student thinks he or she has much to be proud of). We calculate an unweighted average of these questions to build an index.
Column 1 of Table 17 seems to show that school closure is associated with overall higher student self-esteem, yet some individual components of the index show a positive sign (columns 2, 3, 4, 5, 7) and others show a negative sign (columns 6, 9) while still others show no effect (columns 8, 10, 11). It is thus difficult to conclude that self-esteem is unambiguously affected in a single direction.

Table 17: Self-Esteem

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Index	Satisfied	no good	qualities	able	not proud	useless	worth	more respect	failure	positive
			(inverse)			(inverse)	(inverse)		(inverse)	(inverse)	
Math	-0.03	-0.04	-0.04	-0.02	0.01	-0.08	-0.00	-0.03	-0.03	-0.08	-0.07*
	(0.02)	(0.06)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)	(0.04)	(0.03)	(0.05)	(0.04)
Math * School Closed	-0.00	0.10	0.03	0.02	-0.05	0.03	-0.01	-0.12	0.07	0.01	-0.00
	(0.05)	(0.13)	(0.13)	(0.12)	(0.12)	(0.12)	(0.13)	(0.08)	(0.10)	(0.08)	(0.09)
School Closed	0.12***	0.83***	0.45***	0.18*	0.31***	-0.79***	0.23**	0.02	-0.20**	0.02	0.01
	(0.04)	(0.10)	(0.10)	(0.10)	(0.10)	(0.11)	(0.10)	(0.09)	(0.10)	(0.09)	(0.09)
English	0.02	0.02	0.02	-0.00	0.04	-0.04	0.12*	-0.02	-0.07	0.05	0.02
	(0.03)	(0.06)	(0.07)	(0.06)	(0.05)	(0.06)	(0.06)	(0.05)	(0.04)	(0.07)	(0.05)
Observations	821	820	821	815	821	804	821	821	814	816	820
R-squared	0.135	0.286	0.114	0.105	0.116	0.388	0.068	0.084	0.195	0.049	0.122
Mean Dep. Var	2.964	3.160	3.001	3.226	3.185	2.515	3.076	3.201	1.905	3.094	3.262
SD	0.263	0.633	0.608	0.463	0.522	0.674	0.524	0.439	0.406	0.526	0.467

Note: Standard errors clustered at the student level in parentheses.. *** p<0.01, ** p<0.05, * p<0.1. Column 1 shows the estimation of equation 1 with the unweighted average of the self-esteem index as the dependent variable. Columns 2 and 3 show the results of the same regression specification but with each individual

component of the self-esteem index as the dependent variable. The components all range from 1 to 4, with 4 being strongly agree and 1 strongly disagree. They include: “I am satisfied with oneself”, “I think I am no good at all”, “I feel that I have a number of good qualities”, “I am able to do things as well as most others”, “I feel I do not have much to be proud of”, “I certainly feel useless at times”, “I feel that I am a person of worth”, “I wish I could have more respect for myself”, “I sometimes feel that I'm a failure”, and “I take a positive attitude toward myself”. Because columns 3, 6, 7, 9, and 10 ask questions where the ideal response is the lowest possible value, we reverse the response values (i.e. a response of 5 out of 5 for the question in column 6 now indicates that a student thinks he or she has much to be proud of). All regressions control for a student's age, gender, and current year of schooling, as well as the baseline survey response of the dependent variable and a dummy variable “Baseline Missing” equal to 1 if the baseline survey response is missing. If baseline survey response is missing, it is replaced by the value 0 and the dummy variable “Baseline Missing” takes the value 1.

G.0.5 Perceptions on Canada and Kenya

Finally, we present the results for the modules related to perceptions on Canada. The statement is: “Canada is a great place to live” and “Canada is a great place to be”. The responses range from 1 to 4 for the first question, with 1 being strongly disagree and 4 being strongly agree, and 1 to 5 for the second question. In order to avoid the second question weighing more than the first due to a larger scale of possible responses, we rescale each question to range from 0 to 1. We build an index which is the unweighted average of these two questions.

The results are unclear in Table 18. The same analysis for Kenya in Table 19 tends to show that students disagreed more on average with the idea that Kenya is a great place to live during the period in which schools were closed.

Table 18: Canada

	(1)	(2)	(3)
	Index	Canada great place to live	Canada very good place to be
Math	-0.06***	-0.06	-0.26**
	(0.02)	(0.12)	(0.12)
Math * School Closed	0.02	0.60*	-0.46*
	(0.06)	(0.31)	(0.26)
School Closed	0.01	-0.85***	0.17
	(0.04)	(0.21)	(0.15)
English	0.00	-0.09	0.11
	(0.01)	(0.07)	(0.08)
Observations	821	736	820
R-squared	0.097	0.170	0.106
Mean Dep. Var	0.921	3.268	4.522

SD	0.148	0.747	0.824

Note: Standard errors clustered at the student level in parentheses.. *** p<0.01, ** p<0.05, * p<0.1. The table shows the estimation of equation 1. In column 1, the dependent variable is the unweighted average of the index on Canada. Columns 2 and 3 represent the components of this index. In column 2, the dependent variable represents the responses of students to the statement "Canada is a great place to live." The responses range from 1 to 4, with 1 being strongly disagree and 4 being strongly agree. Likewise, the dependent variable in column 3 is the response of students to the statement "Canada is a great place to be." The responses range from 1 to 5, with 1 being "strongly disagree" and 5 being "strongly agree".

Table 19: Kenya

	(1)	(2)	(3)
	Index	Kenya great	Kenya very good
		place to live	place to be
Math	-0.02	-0.15	-0.03
	(0.01)	(0.10)	(0.13)
Math * School Closed	0.03	-0.27	0.43
	(0.03)	(0.20)	(0.27)
School Closed	-0.08***	0.09	-0.63***
	(0.02)	(0.12)	(0.22)
English	0.02	0.11*	0.01
	(0.01)	(0.06)	(0.11)
Observations	821	822	817
R-squared	0.174	0.139	0.143
Mean Dep. Var	0.857	3.575	3.998
SD	0.122	0.644	0.994

Note: Standard errors clustered at the student level in parentheses. *** p<0.01, ** p<0.05, * p<0.1. The table shows the estimation of equation 1. In column 1, the dependent variable is the unweighted average of the index on Kenya. Columns 2 and 3 represent the components of this index. In column 2, the dependent variable represents the responses of students to the statement "Kenya is a great place to live." The responses range from 1 to 4, with 1 being strongly disagree and 4 being strongly agree. Likewise, the dependent variable in column 3 is the response of students to the statement "Kenya is a great place to be." The responses range from 1 to 5, with 1 being "strongly disagree" and 5 being "strongly agree".

H Appendix H: Validity and Reliability of Psychometric Tests

Below we present the tests of internal reliability, test-retest reliability, convergent validity, divergent validity and predictive validity.
Starting with internal reliability, Table 20 displays in column (1) the Alpha Cronbach test of each psychometric scales. The Alpha of the Oral Comprehension, Cross-Cultural Communication, Computer Proficiency, Liking School, Liking Courses, Academic Motivation, and Self-Esteem are all above 0.6, as per the NIH guidelines.[21]
This indicates that the items composing a scale are well correlated with each other; i.e., when participants give a high response for one of the items, they are also likely to provide high responses for the other items. The Aspirations scale has a slightly lower alpha (0.51), not far from the guideline of 0.6. The alpha for the scales Thoughts on Canada and Thoughts on Kenya are 0.41 and 0.40. These scales are less central to the analysis, with only a remote link between tutoring and thoughts on Canada and Kenya; indeed we did not detect any meaningful treatment effect for these scales.

Table 20: Internal Reliability

	(1)	(2)
	Alpha Cronbach	ICC
Oral Comprehension	0.96	0.48**
Cross-Cultural Communication	0.68	0.28**
Computer Proficiency	0.87	0.34**
Aspirations	0.51	0.33**
Liking School	0.70	0.17**
Liking Courses	0.70	0.21**
Academic Motivation	0.64	0.45**
Self-Esteem	0.71	0.28**
Thoughts on Canada	0.41	0.02**
Thoughts on Kenya	0.40	0.28**

For test-retest reliability, we calculate the correlation between repeated waves for the same students. The intraclass correlation is displayed in Column (2). It is above 0.3 for most scales, and always statistically significant at 5 percent.
Convergent validity states that scales measuring the same concepts should positively correlate with each other. In Table 21, we measure the correlation between Oral Comprehension, Cross-Cultural Communication, Computer Proficiency, Aspirations, Liking School, Liking Courses, Academic Motivation and Self-Esteem. The basic intuition is that these scales should be

[21] https://www.ncbi.nlm.nih.gov/books/NBK581902/

positively correlated, e.g., students motivated in class should also like courses. Indeed we find a positive correlation between all these scales, as displayed in the table. The only exception is the correlation between aspirations and proficiency with computer, which is negative. One may argue that these two concepts are not obviously connected, therefore a low correlation may be expected. In other words, one can be good at computers and have low aspirations, or vice versa.

Table 21: Convergent Validity

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Oral Comp.	Cross-Cultural Communication	Computer Proficiency	Aspirations	Liking School	Liking Courses	Academic Motivation	Self Esteem
Oral Comprehension	1							
Cross-Cultural Communication	0.39**	1						
Computer Proficiency	0.31**	0.44**	1					
Aspirations	0.09**	0.06**	-0.07**	1				
Liking School	0.14**	0.25**	0.20**	0.11**	1			
Liking Courses	0.19**	0.27**	0.29**	0.10**	0.95**	1		
Academic Motivation	0.20**	0.21**	0.30**	0.20**	0.45**	0.50**	1	
Self-Esteem	0.14**	0.16**	0.03	0.11**	0.25**	0.22**	0.38**	1

Divergent validity states that there should be no correlation between measures that should not have a relationship. To test for divergent validity, we focus on the scales Thoughts on Canada and Thoughts on Kenya. These scales ask about perceptions on the countries of Canada and Kenya (e.g., Canada is a great place to live; Kenya is a great place to live). These scales should not be connected with academic motivation or liking school; and indeed they are not, as Table 22 shows.

Table 22: Divergent Validity

	(1)	(2)
	Thoughts on Canada	Thoughts on Kenya
Oral communication	0.0666*	-0.1520*
Cross-cultural communication	-0.0084	-0.0623*
Proficiency with computer	0.0011	-0.1810*
Aspirations	0.1235*	0.1108*
Liking School	0.0136	0.0606*

Liking courses	0.0385	-0.0176
Academic Motivation	0.0431	-0.0599
Self Esteem	-0.0078	0.0764*

Finally, predictive validity evaluates how well a scale predicts an outcome. We use grades in school as an outcome in Table 23 below, and find that indeed the psychometric scales of Oral communication, Cross-cultural communication, Liking School, Liking courses, Academic Motivation, and Self Esteem are positively correlated with grades in school. The scale Aspirations is positively correlated with grades, very close to being significant. The scale Proficiency with computer is not correlated with grades, which may be expected since these are different skills. One can be proficient with using a computer, sending emails, but this does not necessarily correlate with grades.

Table 23: Predictive Validity

	(1)
	Grade Total
Oral communication	0.2189*
Cross-cultural communication	0.0801*
Proficiency with computer	-0.0059
Aspirations	0.0635
Liking School	0.0906*
Liking courses	0.0982*
Academic Motivation	0.1275*
Self Esteem	0.0703*

Overall, we find that the psychometric scales used in this paper display internal reliability, test-retest reliability, convergent validity, divergent validity and predictive validity.

I Appendix I: Estimating the Learning Loss

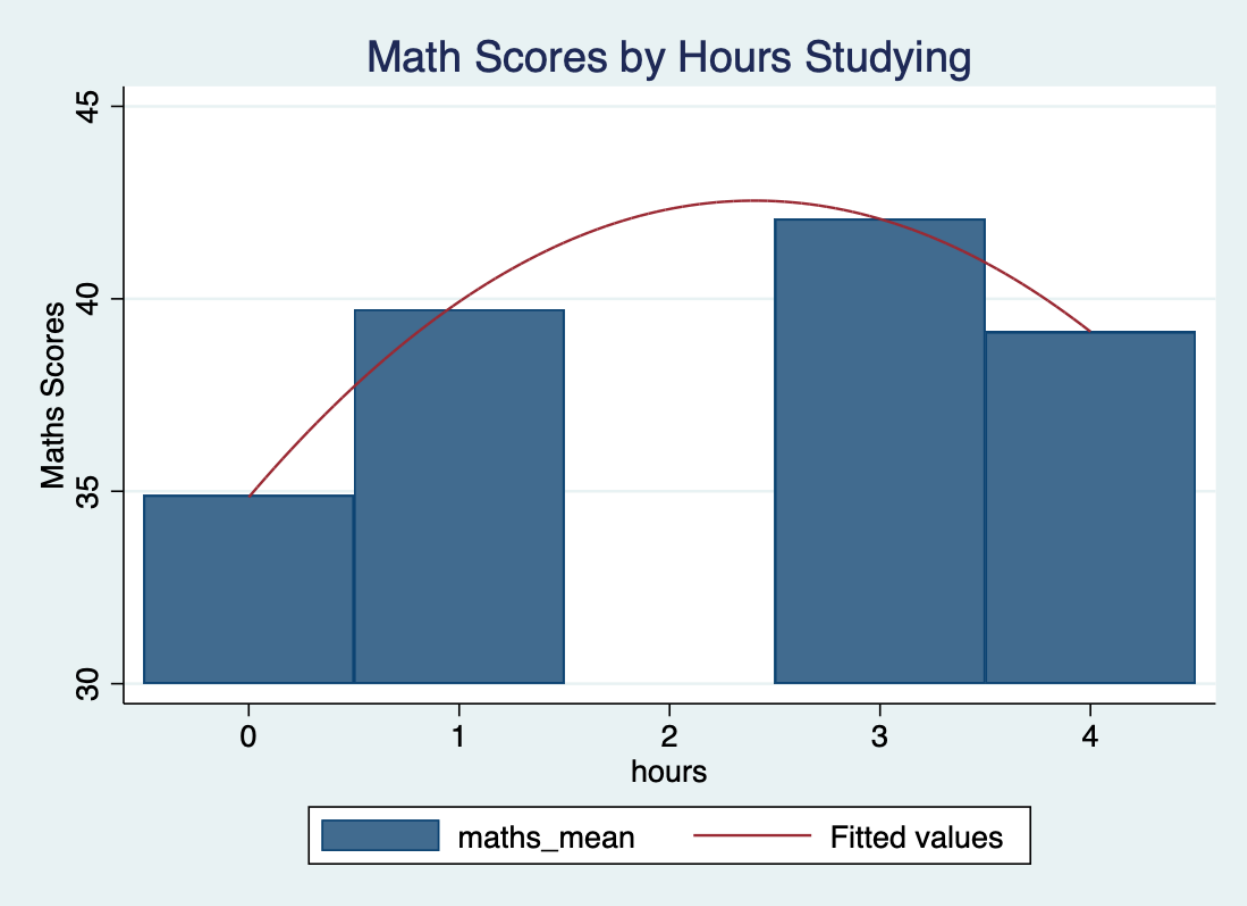
We found that the online tutoring increases grades in Math when the schools are closed, but not when they are open. We use this fact to propose a methodology to estimate the learning loss, other than with a difference-in-differences.

We can summarize the three variables Math , SchoolClosed , and Math*SchoolClosed into one single variable: the number of hours spent studying mathematics. When Math=0 and SchoolClosed=0 , the student is in the control group and the schools are open. In that case, the students receives 3 hours of Math per week (which is the regular teaching load in Math for

grade 6 students in Kenya). When Math=1 and SchoolClosed=0 , the student is treated and receives an additional hour of Math per week.[22]
When Math=0 and SchoolClosed=1 , the student receives no intervention and the schools are closed, such that the student receives no education in Math. Finally, when Math=1 and SchoolClosed=1 , the student receives an hour of Math per week (intervention and schools closed).
Therefore, we construct a variable MathHours_itk as the total number of hours per week student i spends studying mathematics from schooling and tutoring. According to the logic above, it is equal to 3 for the control group when the schools are open, 4 for the treatment group when the schools are open, 0 for the control group when the schools are closed, and 1 for the treatment group when the schools are closed.
Figure 3 below already lets on the idea of decreasing returns to math hours. We find a treatment effect when the schools are closed (for the first hour of Math taught) and no effect when the schools are open (moving from 3 to 4 hours of math, in fact a slightly negative effect but not significant). We superimpose a quadratic fitted line that clearly shows decreasing returns.

Figure 3: Math Hours

[22] One hour of tutoring may not be exactly comparable to one hour of teaching in class. One hour of tutoring may be more than one hour in class since the tutor is teaching one on one as opposed to the teacher teaching to an entire class. One hour of tutoring may be less if there are small interruptions or departures from the tutoring, such as when the tutor tries to get to know the tutee better through regular conversation, or occasional issues regarding the video quality. In any case, the relevant comparison in our analysis is the effect on Math grades with one hour of tutoring after 3 hours of teaching (when the schools are open) and after zero hours of teaching (when the schools are closed). That extra hour of tutoring is comparable. We repeated the analysis assuming that an extra hour of tutoring was equivalent to more or less of an hour in class, and find very similar results.



To capture these decreasing returns, we regress the math grade on Math hours and its squared term in the following specification:

$$y\_itk = \beta 1\ MathHours\_itk + \beta 2\ MathHours\_itk\ 2 + \beta 3\ English\_it + \beta 4\ BaselineMathGrade\_it0 + \beta 5\ BaselineMissing\_itk + \beta 6\ X\_it + \beta 7\ \delta\_tk + \varepsilon\_itk$$

where y_itk represents student i 's math grade in year t , wave k . MathHours_itk is the total number of hours per week student i spends studying mathematics from schooling and tutoring. The squared term of MathHours_itk is also included to capture decreasing returns in a simple way. All specifications are augmented with wave-year fixed effects δ_tk . The identification strategy is that the variation in MathHours_itk is exogenous and provided by the randomized experiment implemented at two different point in time.

Table 24 shows the results below. Column 1 shows the results for the simple regression of math scores onto hours of studying mathematics. One additional hour of studying leads to a higher math score by 7.54 points, or (7.54/13.77=) 0.55 standard deviations. In Table 25 in Appendix I, we add controls in a similar fashion to those in Table 3; that is, we respectively control for the total baseline grade (without math), student characteristics, and all baseline index surveys, and find very similar results.

Table 24: Hours Worked

	(1)
	Math grade
Math Hours	7.54***
	(1.43)
Math Hours Squared	-1.16***
	(0.33)
Wave*Year fixed effects	YES
Controls:	
Baseline Total grade	NO
Age, Gender, School Year	NO
Baseline Survey	NO
Observations	2,170
R-squared	0.355
Mean Dep. Var.	40.82
SD	13.77

This function represents a production function of grades, estimated through a randomized intervention implemented at two different time periods: when schools are open and when schools are closed.

In fact, these results allow us to quantify the learning loss. In regular times, the math grade is 7.54*3-1.16* 3 2 , whereas during the pandemic, the math grade is 7.54*0-1.16*0 , therefore the learning loss is the difference between these two numbers: 12.18.

This estimate is very close to the standard difference-in-difference estimator (we had found -11.95 for the coefficient of SchoolClosed in Table 3), yet it does not rely on the parallel trends assumption. Instead, our estimator relies on a randomized experiment, implemented at two different points in time, such that we can evaluate the decreasing returns to hours of teaching in math in a production function of grades. The fact that these two methodologies yield relatively similar estimates support the claim that school closures causally created a large learning loss.

We corroborate the evidence that we provided earlier for the diminishing marginal returns of hours studying on math grades in Figure 3 as well as Table 24 by adding additional controls. More specifically, we augment the specified model with baseline total grade, age, gender, school year, and baseline survey responses. In all specifications, the number of hours spent studying math is statistically significant and positive, while the hours squared term is statistically significant and negative, albeit with a coefficient of much lower magnitude. This confirms the trend of diminishing marginal returns highlighted in Figure 3.

Table 25: Hours Worked

	(1)	(2)	(3)
	Dependent Variable: Math Grade		
Hours	6.93***	7.14***	7.20***
	(1.45)	(1.50)	(2.10)
Hours Squared	-1.09***	-1.09***	-1.10***
	(0.31)	(0.31)	(0.37)
Wave*Year fixed effects	YES	YES	YES
Controls:			
Baseline Total Grade	YES	YES	YES
Age, Gender, School Year	NO	YES	YES
Baseline Survey	NO	NO	YES
Observations	2,170	2,170	2,170
R-squared	0.393	0.399	0.431
Mean Dep. Var.	40.82	40.82	40.82
SD	13.77	13.77	13.77

Note: Standard errors clustered at the student level in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Column 1 shows the results while controlling for the baseline total grade, with a student's math grade as the dependent variable. Column 2 augments this specification by controlling for a student's age, gender, and the year of schooling they are currently completing. Column 3 adds to column 2 by including in the list of controls the baseline averages of various indices from the survey data. These indices include: oral comprehension, computer proficiency, cross-culture communication, motivation, self-esteem, future aspirations, liking school, liking classes, thoughts about Canada, and thoughts about Kenya.

Author Statement

McGill

Department of Economics
McGill University

855 Sherbrooke Street West
Montreal, Quebec, Canada H3A 2T7

Département de sciences économiques
Université McGill

855, rue Sherbrooke ouest
Montréal (Québec) Canada H3A 2T7

Telephone: (514) 398-3030
Facsimile: (514) 398-4938
Web Site: http://www.mcgill.ca/economics/

September 9, 2024

RE: Disclosure statement

Declarations of interest: none

Sincerely,

Matthieu Chemin
Associate Professor
Department of Economics
McGill University
https://www.matthieuchemin.com/
https://elimu.lab.mcgill.ca/