Impacts of an Information and Communication Technology-Assisted Program on Attitudes and English Communication Abilities: An Experiment in a Japanese High School

Yuki Higuchi, Miyuki Sasaki, and Makiko Nakamuro*

We conducted a randomized experiment targeting 322 Japanese high school students to examine the impacts of a newly developed English-language learning program. The treated students were offered an opportunity to communicate for 25 minutes with English-speaking Filipino teachers via Skype several times a week over a 5-month period as an extracurricular activity. The results show that the Skype program increased the interest of the treated students in an international vocation and in foreign affairs. However, the students did not improve their English communication abilities, as measured by standardized tests, probably because of the program's low utilization rate. Further investigation showed that the utilization rate was particularly low among students demonstrating a tendency to procrastinate. These results suggest the importance of maintaining students' motivation to keep using such information and communication technology-assisted learning programs if they are not already incorporated into the existing curriculum. Having procrastinators self-regulate may be especially crucial.

Keywords: Japan, learning English, procrastination, randomized controlled trial, Skype
JEL codes: C93, H40, I21

I. Introduction

Providing students with high-quality learning resources is critically important in improving the quality of education. In recent years, information and

*Yuki Higuchi (corresponding author): Faculty of Economics, Sophia University, Japan. E-mail: higuchi@sophia.ac.jp; Miyuki Sasaki: Faculty of Education and Integrated Arts and Sciences, Waseda University, Japan. E-mail: miyuki.sasaki@waseda.jp; Makiko Nakamuro: Faculty of Policy Management, Keio University, Japan. E-mail: makikon@sfc.keio.ac.jp. This study was conducted as a part of the Measurement of the Qualities of Health and Education Services, and Analysis of their Determinants project undertaken at the Research Institute of Economy, Trade and Industry. We would like to thank Tomohiko Inui, Yukichi Mano, Ryoji Matsuoka, Shinpei Sano, an anonymous referee, and participants of the Asian Development Bank–International Economic Association Roundtable for helpful comments and suggestions. We also acknowledge Takeshi Kamimura, Tomohisa Kato, and Tomoya Sugiyama for their active research collaboration. This research was financially supported by MEXT/JSPS KAKENHI Grant Number: 18H05314, Grant-in-Aid for Research at Nagoya City University, where the first and second authors were affiliated with until March 2020, and Keio University. All errors are our own. The usual ADB disclaimer applies.

Asian Development Review, vol. 37, no. 2, pp. 100–133
https://doi.org/10.1162/adev_a_00151

communication technology (ICT) has increasingly been used as an alternative to more conventional resources (e.g., Gee and Hayes 2011, Levy 2009). Such ICT-assisted educational resources can be best used to help overcome the limitations of conventional resources. In particular, because ICT can provide customized and self-paced learning opportunities, the use of ICT in education has huge potential to improve the effectiveness of home learning.

According to surveys by Bulman and Fairlie (2016) and Snilstveit et al. (2016), the classroom use of ICT generally has positive impacts, especially for students in lower grades studying math or science. While earlier observational studies found large positive impacts of home use of ICT on students' academic outcomes, these studies suffered from the selection bias that students or teachers with unobserved high ability or motivation tended to introduce the new ICT-assisted resources. More recent experimental studies tended to find smaller or even no impacts.[1] Such mixed results for the home use of ICT partly reflect differences in the grades of the sampled students, their proficiency levels, sampled countries, and studied or targeted subjects; however, we particularly need evidence on whether the home use of ICT can compensate for the weaknesses of conventional education resources.

To test the usefulness of the home use of ICT in complementing current education programs, we conducted a randomized controlled trial (RCT) that provided ICT-assisted resources for Japanese high school students learning English. In contrast to the high internationally normed performance of Japanese students in reading, math, and science—as measured by the Organisation for Economic Co-operation and Development's Program for International Student Assessment for Grade 9 students—their performance in English has been far from satisfactory. According to a nationwide English test conducted in 2014 by the Ministry of Education, Culture, Sports, Science and Technology, Japan (MEXT), a majority of Grade 12 students ranked at the lowest level (A1) in the Common European Framework of Reference for Languages, with their speaking performance lowest among the four skills measured. Based on these results, MEXT recognized that the quality of English education, particularly in nurturing speaking ability, should be improved (MEXT 2015a). As conventional English education programs in Japan have been unsuccessful, there is scope for the use of ICT-assisted resources to improve the quality of such education.

We experimentally introduced a newly developed online English learning program as an extracurricular activity to 322 Japanese students in Grade 10. This online program is an individualized, self-paced program in which students communicate with English-speaking Filipino interlocutors, mostly consisting of

[1] This is reminiscent of Glewwe et al. (2004), who compared an observational study with an experimental one and found that the large positive impact of the introduction of flipcharts to Kenyan schools found in the observational study was no longer detected in the experimental one.

current students or graduates of the University of the Philippines, the top national university in the country. The students can communicate with them at mutually convenient times via Skype using learning materials of their own choice. This program is an example of human resource arbitrage from developing to developed countries with the help of modern ICT technology. Although it is beyond the scope of this paper, the program may have positive impacts not only on the Japanese-student side but also on the Filipino-instructor side by creating earning opportunities.

We introduced the Skype English program with a crossover design.[2] First, we randomly selected half of our sample (161 students) to be given the opportunity to use the program for 5 months from July to November 2015, while the remaining 161 students were given the opportunity to use the program for 5 months from January to May 2016. While all the students had an equal opportunity to use the program by May 2016, only half of them had taken this opportunity as of December 2015, when we conducted the endline survey. We therefore refer to the students exposed to the program in the first round (July–November 2015) as the treatment group and those exposed to it in the second round (January–May 2016) as the control group.[3]

Combining program usage records and panel data collected before and after the introduction of the program to the treatment group (but not yet to the control group), two main findings emerge. First, the program changed the attitudes of the treated students positively, especially in terms of their interest in an international vocation and in foreign affairs. In particular, our estimates of the local average treatment effect (LATE) suggest that the effects were large for students with greater program utilization. This finding is important because past longitudinal studies suggested that it is difficult to change students' attitudes toward an international vocation and foreign affairs when they study a foreign language (Ortega and Iberri-Shea 2005). This may be particularly the case in the Japanese school environment, which is known to have a monocultural and monolingual orientation. Furthermore, Sasaki (2011); Yashima (2002); and Yashima, Zenuk-Nishide, and Shimizu (2004) found that such attitudinal change among Japanese students will eventually lead to improvements in their English communication skills.

Second, despite the positive impacts on the students' attitudes, there is no measured impact on their English communication skills. This may be attributed

[2] Although an RCT is now recognized as best practice in impact evaluation, it is extremely difficult to run such a trial in Japanese public schools, where priority is given to equality of resource allocation within the same cohort of students. Hence, as a second-best strategy, we conducted an RCT with a crossover design, ensuring that all students received the same treatment within the same academic year, with the only difference being in respect to the timing of the treatment. A shortcoming of this strategy is that the evaluation period is less than 6 months, but we emphasize that our study is a unique RCT conducted in a public school in Japan.

[3] A referee suggested to additionally use a difference-in-differences (DiD) "in reverse" approach, exploiting the change in status of the control group from before-treatment to after-treatment, while the treatment group remained after-treatment status (Kim and Lee 2019). We, however, were unaware of this approach and did not conduct a survey or a standardized English test after the intervention with the control group. We note that DiD "in reverse" is a useful approach in a crossover RCT in general.

to the low intensity of the program (25 minutes per lesson) in comparison with the students' concurrent regular English classes (50 minutes per lesson on most weekdays) as well as the program's low utilization rate. Only 10 of the 161 students in the treatment group took 50 or more lessons over the 5-month period, as recommended by the program provider, and 31 students took no lessons over the same period. In addition, regression analyses show that the utilization rate was particularly low among students with a tendency to procrastinate, which is consistent with the emerging literature on self-control problems (e.g., Duckworth, Milkman, and Laibson 2018). These findings warrant further research on how to improve and maintain students' motivation, particularly those with a tendency to procrastinate, to adopt home-use ICT programs such as the one targeted in this study.

The remainder of this paper is organized as follows. Section II describes our experiment, including the sample, timeline, and details of the intervention. Section III discusses sample balance and program utilization, and section IV presents the estimated program impacts. Finally, section V contains a summary of the findings and implications for future studies.

II. Experiment

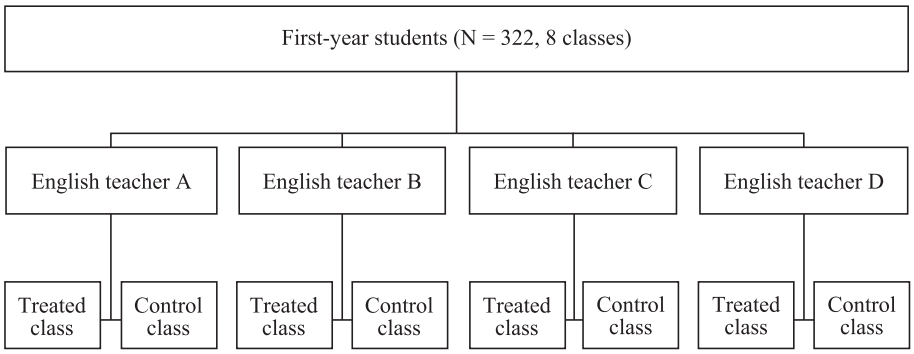
A. Sample

We collaborated with a public high school that is a top-tier school in central Japan. This school was selected by the Government of Japan in 2015 as one of the 112 Super Global High Schools among the 4,939 high schools in Japan. Super Global High Schools receive extra budgetary support to nurture globalized leaders with high levels of interest in societal problems, communication skills, and problem-solving abilities, who will play internationally active roles in the future (MEXT 2015b). The school agreed to introduce the online program as an extracurricular activity.

Our sample consisted of all 322 first-year high school students (Grade 10) who were newly admitted to the school a few months before the experiment.[4] In Japan, high school admissions, whether public or private, are mostly based on students' academic performance on the entrance examination, with students subsequently tracked into different high schools of varying quality. After our sample students were admitted to our target high school, they were randomly assigned to one of eight classrooms, each consisting of 40 or 41 students. Classroom assignment

[4] We provided all the parents of the sample students with information on our research and its purpose before commencing data collection and intervention. As the parents of one student refused to provide data for our analyses, we excluded the data collected from that student. Thus, the sample size is 321 in our empirical analyses.

Figure 1. **Randomization**



Source: Authors' illustration.

was not affected by any preexisting peer groups; we took advantage of this to attain randomization in our experiment.

Further, each of the four full-time English teachers in the school were randomly assigned to teach two of these eight classes. To achieve balance in the quality of the English teachers in the classroom, we stratified the sample of students at the teacher–classroom level, randomly assigning one of the two classes instructed by each English teacher to the treatment group and the other to the control group (Figure 1). In sum, we have four treatment classes (with 160 students) and four control classes (with 161 students). Although our experiment may suffer from a small number of clusters (i.e., eight classes), the classroom-level intracluster correlation coefficients for outcome variables at the baseline survey are close to 0, indicating that there is little correlation of responses within a cluster, and thus, our randomization can be considered as being close to the student-level randomization.[5]

B. Timeline

Before introducing the program, we conducted a baseline survey designed to collect information on the students' characteristics and attitudes toward English communication. The survey was conducted in June 2015, using a mark-sheet questionnaire we developed. The timeline of our research is presented in Table 1.

Soon after the baseline survey, the sample students took the Versant speaking test (Pearson Inc. 2008), a standardized test designed to evaluate the oral English

[5]The classroom-level randomization will help us mitigate the violation of the Stable Unit Treatment Value Assumption caused by spillover effects among students in the same classroom. While admitting that it is technically difficult to separate the direct effect of our intervention from the indirect effect through their peers in the classroom-level randomization, as pointed out by Imbens and Wooldridge (2009), we think that the degree of such indirect effect is limited because our outcome variables are individual measures of attitudes and test scores, which are more likely to be affected by interactions with English teachers than by those with their classmates.

Table 1. **Research Timeline**

June 2015	Baseline survey and (i) Versant test
1 July 2015	Online English program for the treatment group starts
July 2015	(ii) Benesse test
November 2015	(ii) Benesse test and (iii) GTEC English test
30 November 2015	Online English program for the treatment group ends
December 2015	Endline survey and (i) Versant test
January–May 2016	Online English program for the control group

GTEC = Global Test of English Communication.
Source: Authors' compilation.

skills (integrated listening and speaking) of nonnative English speakers.[6] The test was administrated solely for this research project (although the results were shared with the students as feedback) to construct our measure of English communication ability. Following the survey and the Versant test administered in June 2015, we commenced the intervention on 1 July 2015. The students in the treatment classes were provided with opportunities to use the online program free of charge, although the market price of the program was ¥5,800 (about $52) per month. This included one 25-minute lesson for every day of the intervention period.

Soon after our intervention commenced, the students took a nationally administrated English test developed and distributed by Benesse Co. The test is a mock university entrance exam designed primarily to measure students' English reading ability. The sample students took a similar test again in November, toward the end of our intervention. Although the tests were not taken for the purpose of our study, the school agreed to share the results with us to be used as another measure of the students' English abilities. In addition, in November, the students took the Global Test of English Communication (GTEC), a standardized test developed and distributed by Benesse Co. to evaluate reading, listening, writing, and speaking skills in English.[7] The school also agreed to share the results of this test with us.

In December 2015, when only the treated students had received the program, we conducted an endline survey and Versant test. In other words, to investigate the effects of the online program, the treatment and the control classes were compared using a difference-in-differences (DiD) design. To mitigate inequality between the two groups (as mentioned above), we provided the same amount of intervention

[6] We chose this particular test because of its reported high validity and reliability among populations similar to the sample in the present study and because it requires a relatively short time (20 minutes) to conduct compared with other English communication tests (e.g., TOEFL iBT). During the Versant test, the students listened to questions spoken in English and provided verbal answers in English. Their answers were recorded and automatically marked online. The test was conducted by class in a computer room inside the school, and thus, the test-taking environment was essentially the same for all students. The Versant test scores ranged from 20 to 80 and involved four criteria: (i) sentence mastery, (ii) vocabulary, (iii) fluency, and (iv) pronunciation. The scores correspond with the levels of the Common European Framework of Reference for Languages: for example, a Versant score of 20–25 is equivalent to the lowest (A1) level, while a score of 79–80 is equivalent to the highest (C2) level.

[7] The test consists of 30 multiple-choice reading items (24 minutes), 30 multiple-choice listening items (13 minutes), 3 performative writing items (26 minutes), and 4 performative speaking items (12 minutes).

with a time lag, with the program being made available to the control classrooms from January to May 2016. By the end of May 2016, all 322 students had been exposed to the same intensity of intervention (or lack thereof).

C. Intervention

Our intervention consisted of providing the sampled students with opportunities to use the online program. In contrast to conventional face-to-face English learning methods, in this program, learners and teachers do not have to be present in the same space. In addition, learners can be matched with teachers on a more flexible basis because learners can select among available teachers at a time of their convenience. Such online English programs have become increasingly popular among Japanese businesspeople, partly because of time flexibility advantages and partly because of the low cost of such programs relative to similar face-to-face English learning programs offered by commercial conversation or cramming schools. However, according to the baseline survey that we conducted before the beginning of the intervention, 65% of the students had never heard of this type of online English learning program, only one student was using such a program, and another 10 had used one in the past. In this baseline survey, 30% of the students responded that they would be very willing to use the program if given the opportunity, and another 50% responded that they were moderately willing to use it. Hence, while the program was new to most of the students, it was favorably perceived at the beginning of our intervention.

The online program was provided to the students outside of their regular English classes. Each lesson took 25 minutes, and the students were recommended to take one lesson every 3 days (i.e., 10 lessons a month, or 50 in total) to take full advantage of the program. The students could make an appointment for a lesson at any time between 6 a.m. and 1 a.m. on the following day and could choose any of the available teachers. If the student's preferred teacher was not available at the time of their convenience, they were able to choose another time slot or another available teacher in the same slot. The pool of teachers consisted mostly of current students or graduates of the University of the Philippines. Because English is the language of instruction in their home university and also because they were screened on the basis of the company's strict hiring criteria, we judged that the quality of the teachers was reasonably guaranteed. While some of the teachers spoke Japanese, participating students had to communicate entirely in English with the help of the chat (texting) function in Skype. Students were free to choose appropriate study materials for each lesson from a wide range of materials provided by the program, including daily conversation, academic talk, grammar and vocabulary, and business English. In other words, the participants' choice of teachers, time slots, and study materials were their decision entirely. Most importantly, while we provided the students with opportunities to use the program at home, it was ultimately up to them whether and

Table 2. **Balance Check**

	Treatment		Control		Difference *p*-value for Equality of Means
	Mean	N	Mean	N	
Procrastination (*z*-score)	−0.020	157	0.021	155	0.72
Male (1 = yes)	0.50	159	0.50	161	0.99
English since Grade 1 or 2 (1 = yes)	0.42	156	0.40	154	0.63
English since Grade 3 or 4 (1 = yes)	0.41	156	0.42	154	0.92
English since Grade 5 or later (1 = yes)	0.17	156	0.19	154	0.62
Been abroad (1 = yes)	0.39	157	0.37	159	0.66
Own room (1 = yes)	0.89	157	0.84	159	0.15
Own personal computer (1 = yes)	0.08	152	0.12	154	0.20
Own tablet (1 = yes)	0.23	156	0.16	159	0.10
Commuting 20 minutes or less (1 = yes)	0.26	156	0.21	155	0.24
Commuting 21–40 minutes (1 = yes)	0.38	156	0.42	155	0.53
Commuting 41–60 minutes (1 = yes)	0.26	156	0.26	155	0.87
Commuting 61 minutes or more (1 = yes)	0.10	156	0.11	155	0.70
Belongs to sports club (1 = yes)	0.65	156	0.57	155	0.19
Number of books at home[a]	2.66	154	2.33	155	0.06

N = number of observations.
Notes: [a]Number of books at home; 0 = none, 1 = approximately 20, 2 = approximately 50, 3 = approximately 100, 4 = approximately 200, and 5 = over 300.
Source: Authors' calculations.

how often to take the lessons, especially because their participation did not affect their grades.

One of the biggest advantages of this online program is its cost-effectiveness. The government launched the Japan Exchange and Teaching Program in 1987, which involved providing English-speaking aides known as Assistant English Teachers (AETs) to Japanese English teachers in primary, middle, and high schools (Grades 1–12). This program has expanded since then—a total of 5,163 AETs were employed as of 2017. The individual annual cost for an AET is approximately $53,000, including salary, coordination, and transportation, while the market price of this English program is $600 per year. Based on the program provider's back-of-the-envelope calculation, the program enables students to devote 15 times more minutes to speaking with English-speaking partners than speaking with an AET for every dollar spent.

III. Balance and Program Utilization

A. Balance

Table 2 presents the basic characteristics of students that could potentially influence the take-up rate and effects of the online program. As the literature finds that a lack of self-control, including procrastination, can result in poor

test performance or low grades (e.g., Golsteyn, Grönqvist, and Lindahl 2014; Onji and Kikuchi 2011), we constructed an index of procrastination as a control variable based on the six questions to rate students' perception of themselves, taken from Osaka University (2013) and Honda and Nishijima (2007). The questions (originally written in Japanese and translated by the authors) included items such as "Are you a person who postpones plans even when you make them?" and "Are you a person who is happy as long as you are having fun now?" The students answered all six questions with categorical responses: (i) yes, (ii) moderately yes, (iii) 50/50, (iv) moderately no, or (v) no. We assigned a score of 4 to the answer yes, 3 to moderately yes, 2 to 50/50, 1 to moderately no, and 0 to no. We then aggregated the scores for all six questions to construct a single index of procrastination, which ranged from 0 to 24 (maximum of 4 multiplied by 6 items). These aggregated scores were normalized by subtracting the sample mean and then dividing by the standard deviation. The mean z-score of the procrastination index is -0.02 among the treatment group and 0.021 among the control group; importantly, these means are statistically not different.

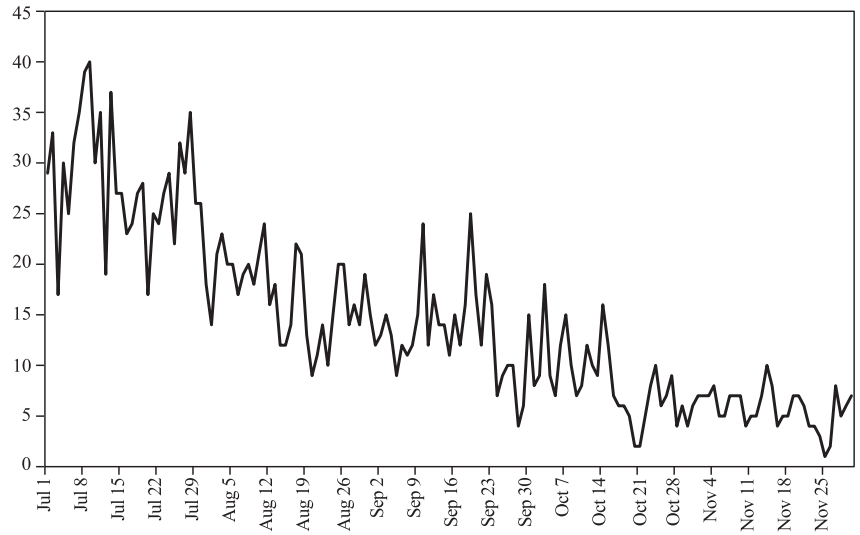
Other control variables include gender, past exposure to English (whether the student has been abroad and the grade at which they started learning English in primary school), and current study environment (having their own room and electronic device, such as a personal computer connected to the internet or a tablet, commuting time to school, and membership of a school sports club), as well as their family background (number of books at home and parental educational attainment).[8] We also collected information on smartphone ownership, but almost all of the students (96%) owned one so we do not include this variable as a control. The differences in means between the two groups are statistically insignificant at the 5% level for all the variables, indicating that randomization was performed successfully.

B. Program Utilization

Figure 2 shows daily changes in the number of students who took the lessons based on program usage records. Of the 160 students assigned to the treatment group, the average number of students who took lessons each day was 25 in July 2015. However, if all students had completed the recommended 10 lessons a month, that number would be 52 (10 lessons multiplied by 160 students and divided by 31 days). Thus, the take-up rate in the first month of the intervention was about 50%. Moreover, the number of students taking lessons decreased gradually, presumably because the novelty effect faded and peer pressure was muted by the summer

[8] As a number of students (27 in the treatment group and 21 in the control group) did not report their parental educational attainment, we do not use the variables of father's education and mother's education. Instead, we use the variable of number of books at home as a proxy of parental socioeconomic status. Kawaguchi (2016) found a correlation between the number of books at home and parents' earnings among Japanese Grade 10 students.

Figure 2. **Daily Change in Number of Students Taking Lessons, 2015**



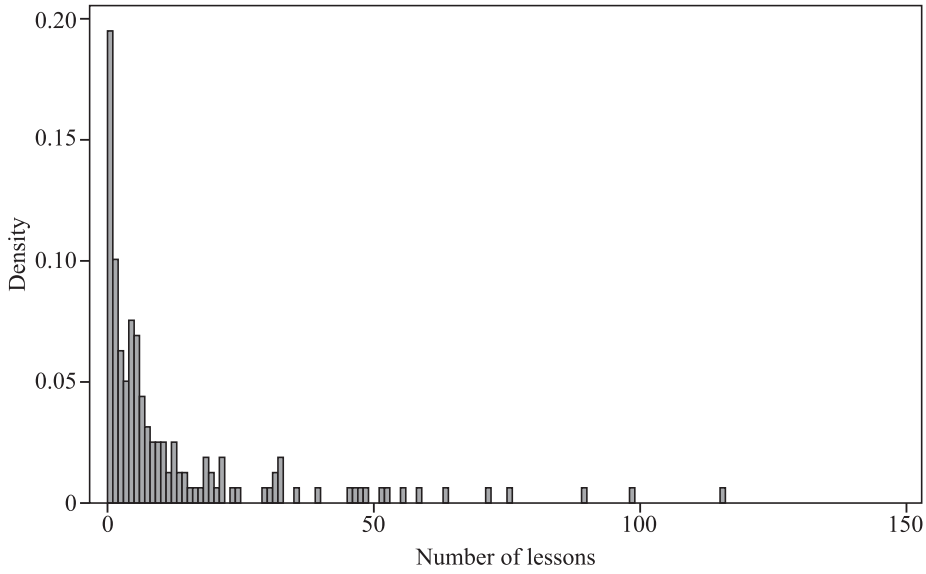
Source: Authors' calculations.

vacation, which started during the last week of July, with the average number falling to 15 in August, 12 in September, 6 in October, and 5 in November. While Figure 2 shows daily changes in program utilization, Figure 3 shows the student-level number of lessons taken during the intervention period. Thirty-one (19%) of the 160 students never took any lessons in the 5-month period, and 57 (36%) took five or fewer lessons. Only 23 students (14%) completed 25 or more lessons, one-half of the recommended number, of whom only 10 (6%) completed the recommended 50 or more lessons.

To identify the factors associated with program utilization, we estimated the ordinary least squares models while controlling for the English teacher dummies. Column 1 shows that the effect of the procrastination index is negative and significant, illustrating the detrimental effect of procrastination on program utilization. The significance of this variable remains robust and consistent, even after the variables listed in Table 2 are controlled (column 2). In terms of the size of the effects, a 1 standard deviation increase in the procrastination index reduces the number of lessons by about 4 times, where the mean was 12.2 times; thus, the influence of procrastination seems nonnegligible.

As the program was new to most of the students and the first few trials of the program are critical for subsequent utilization, we estimated a linear probability model, where the dependent variable is coded as a dummy variable that equals 1 if a student has ever used this Skype program and 0 otherwise. Indeed, according to our informal interviews with some of the students, regular Skype users started to like the program as they proceeded through the initial few talks with Filipino

Figure 3. **Distribution of Lessons Taken by a Student over 5 Months**



Source: Authors' calculations.

interlocutors, whereas nonusers felt hesitant to take the first lesson. Columns 4–6 show the results, and the procrastination variable is negative and significant.

Table 3 also shows that the English teacher dummies are large in magnitude and statistically significant. For instance, a student with English teacher D was about 40 percentage points less likely to have ever used the program than a student with English teacher A (base category). The degree of in-class encouragement and reminders substantially differed from one teacher to another, with teacher A, who is the most senior and experienced among the four teachers, providing more encouragement and more frequent reminders to students to participate in the Skype tasks. According to our informal interviews, this teacher frequently asked the students whether they used the program to put gentle pressure on them as well as to share their experiences with other classmates. This teacher also posted an eye-catching message in the classroom to regularly use the program. These observations suggest that the frequencies of such promotive acts from teachers may be critical to the home use of ICT-assisted inputs.

IV. Impacts

A. Descriptive Analyses: Attitudes

We included two sets of outcome measures to evaluate the impacts of the online program: (i) attitudes and (ii) English communication abilities.

Table 3. **Correlates of Program Utilization (Ordinary Least Squares Estimation)**

	(1)	(2)	(3)	(4)	(5)	(6)
	Number of lessons taken in 5 months			= 1 if completed at least one lesson in 5 months		
Procrastination	−3.82**	−4.35**	−3.88*	−0.097***	−0.085**	−0.084**
[*z*-score]	(−2.60)	(−2.26)	(−1.96)	(−3.26)	(−2.58)	(−2.49)
Male		−1.41	−0.083		−0.12*	−0.12*
(1 = yes)		(−0.29)	(−0.02)		(−1.88)	(−1.81)
English since Grade 3 or 4x		1.09	0.0079		0.052	0.054
(1 = yes)		(0.29)	(0.00)		(0.80)	(0.83)
English since Grade 5 or later		−2.72	−3.23		0.039	0.044
(1 = yes)		(−0.59)	(−0.72)		(0.44)	(0.49)
Been abroad		−0.70	0.11		−0.11*	−0.11*
(1 = yes)		(−0.22)	(0.03)		(−1.67)	(−1.75)
Own room		−3.72	−4.26		−0.15*	−0.15*
(1 = yes)		(−0.71)	(−0.82)		(−1.81)	(−1.79)
Own personal computer		1.17	1.37		−0.15	−0.15
(1 = yes)		(0.21)	(0.25)		(−1.24)	(−1.25)
Own tablet		−0.32	0.38		0.067	0.068
(1 = yes)		(−0.07)	(0.09)		(1.04)	(1.06)
Commuting time 21–40 minutes		6.68*	5.51		0.011	0.013
(1 = yes)		(1.76)	(1.49)		(0.15)	(0.18)
Commuting time 41–60 minutes		4.42	4.40		−0.024	−0.026
(1 = yes)		(0.89)	(0.89)		(−0.27)	(−0.29)
Commuting time 61 minutes		1.37	1.35		0.17	0.16
or over (1 = yes)		(0.30)	(0.29)		(1.49)	(1.37)
Sports club		−1.53	−2.88		−0.12*	−0.12*
(1 = yes)		(−0.33)	(−0.65)		(−1.94)	(−1.82)
Number of books		−0.36	−0.70		0.046**	0.046**
[1–6]		(−0.28)	(−0.58)		(2.43)	(2.39)
Baseline international posture			0.28			0.013
(*z*-score)			(0.20)			(0.44)
English teacher B	−0.14	−0.98	−2.77	−0.21***	−0.24***	−0.24***
(1 = yes)	(−0.03)	(−0.21)	(−0.63)	(−3.21)	(−3.45)	(−3.35)
English teacher C	0.038	−1.31	−1.05	−0.19***	−0.17**	−0.17**
(1 = yes)	(0.01)	(−0.22)	(−0.18)	(−3.18)	(−2.40)	(−2.43)
English teacher D	−7.23**	−10.2**	−9.91**	−0.42***	−0.39***	−0.40***
(1 = yes)	(−2.09)	(−2.37)	(−2.27)	(−5.28)	(−4.97)	(−4.97)
Mean of the outcome variable		12.2			0.81	
R-squared	0.064	0.107	0.099	0.192	0.352	0.352
Adjusted *R*-squared	0.039	−0.002	−0.021	0.170	0.272	0.266
No. of observations	157	147	146	157	147	146

Notes: Estimated coefficients are reported here. ***, **, and * indicate 1%, 5%, and 10% levels of statistical significance, respectively. Numbers in parentheses are *t*-statistics based on heteroscedasticity-robust standard errors. The base category for the English-since variable is "English since Grade 1 or 2," for the commuting time variable it is "Commuting time 20 minutes or less," and for the teacher dummies it is "Teacher A."
Source: Authors' calculations.

To quantitatively measure any changes in students' attitudes toward English communication before and after the intervention, we employed two motivational attributes that have been found to influence students' second-language

development: (i) international posture and (ii) willingness to communicate (WTC) (e.g., Yashima, Zenuk-Nishide, and Shimizu 2004). First, the construct of international posture was operationally defined as a composite of four subconstructs: (i) intercultural orientation; (ii) interest in an international vocation; (iii) reactions to different customs, values, or behaviors; and (iv) interest in foreign affairs. These subcomponents and corresponding items were adapted from those made available on the homepage of Professor Tomoko Yashima, who first introduced this construct to the field of applied linguistics.[9] This construct has proved to be one of the most distinct and significant factors explaining students' motivation, especially in English-as-a-foreign-language contexts (see, for example, Dörnyei and Ryan 2015). Using all 22 available items (seven for subcomponent 1, six for subcomponent 2, five for subcomponent 3, and four for subcomponent 4), we then created questions requiring either yes or no answers. Although the original versions of the 22 questions required responses using a six-point Likert scale, we simplified it to yes–no answers to avoid causing excessive fatigue among the students, who had to respond to many questions in our survey. We computed a score for each of the four subcomponents of international posture and then computed total scores, which ranged from 0 to 22, with a higher score indicating a more internationally oriented student. Finally, we computed *z*-scores for the total score as well as for the four subcomponents.[10]

Panel A of Table 4 presents the means of the international posture scores by group, before and after our intervention with the treatment group (but not yet with the control group). First, the means of all the scores before the intervention were not statistically different between the two groups (see the *p*-values reported on the right). For instance, the baseline mean *z*-score for the treatment group was 0.042, which was slightly higher than the control group mean of −0.041, but the scores are not statistically different. After the intervention, however, the total score became higher among the treatment group than the control group, and the difference is statistically significant at the 5% level. If we examine the subcomponents, a significant difference is observed for subcomponent 2 (interest in an international vocation) and subcomponent 4 (interest in foreign affairs).

Interestingly, the total score dropped from the baseline mean of −0.041 to an endline mean of −0.172 among the control group (*z*-scores were computed using the means and standard deviations among the baseline samples), which is a decline of 0.13 standard deviations. This declining trend was particularly observable for subcomponents 1 and 2, which suggests that the motivation of students to learn English shifted from a more to less internationally oriented one: preparation for university entrance exams. In the top-tier high school where we conducted

[9] Tomoko Yashima. Kokusai. http://www2.ipcku.kansai-u.ac.jp/~yashima/data/kokusai.pdf (accessed April 15, 2019).

[10] Appendix Table A1 presents regression results that analyze the baseline correlates of the international posture *z*-score as well as the baseline correlates of our other outcome variables discussed below.

Table 4. **Differences in Attitudes and English Communication Test Scores by Group**

A. International posture and willingness to communicate

	Treatment		Control		Difference *p*-value for Equality of Means
	Mean	N	Mean	N	
Total international posture [*z*-score, 22 criteria]					
Baseline	0.042	156	−0.041	159	0.47
Endline	0.068	155	−0.172	157	0.05
Sub 1. Intercultural approach tendency [*z*-score, 7 criteria]					
Baseline	0.024	157	−0.024	159	0.67
Endline	−0.091	155	−0.162	157	0.55
Sub 2. Interest in international vocation [*z*-score, 6 criteria]					
Baseline	0.011	157	−0.011	159	0.84
Endline	0.054	155	−0.170	157	0.05
Sub 3. Reaction to different customs [*z*-score, 5 criteria]					
Baseline	0.034	156	−0.033	159	0.56
Endline	0.010	155	−0.031	157	0.71
Sub 4. Interest in foreign affairs [*z*-score, 4 criteria]					
Baseline	0.068	157	−0.067	159	0.23
Endline	0.259	155	−0.076	157	0.01
Willingness to communicate [*z*-score, 8 criteria]					
Baseline	0.063	156	−0.063	155	0.27
Endline	−0.082	155	−0.27	156	0.09
Cambodia study tour (1 = yes)					
Endline	0.101	159	0.068	161	0.30

B. English communication test

	Treatment		Control		Difference *p*-value for Equality of Means
	Mean	N	Mean	N	
(i) Versant score [*z*-score]					
Baseline	0.095	142	−0.093	146	0.11
Endline	0.671	124	0.406	141	0.05
(ii) Benesse score [*z*-score]					
Baseline	−0.032	156	0.031	158	0.58
Endline	−0.030	156	0.030	156	0.60
(iii) GTEC overall score [*z*-score]					
Endline	0.002	158	−0.001	160	0.98
Sub 1. Reading					
Endline	−0.012	159	0.012	161	0.83
Sub 2. Listening					
Endline	0.034	158	−0.033	161	0.54
Sub 3. Writing					
Endline	0.024	158	−0.023	160	0.68
Sub 4. Speaking					
Endline	−0.011	159	0.011	161	0.84

Continued.

Table 4. *Continued.*

GTEC = Global Test of English Communication.
Notes: *z*-scores are computed using the means and standard deviations among the baseline samples for international posture, willingness to communicate, and Versant score. The level of the Benesse test is different from one test to another, as it is in accordance with the school curriculum; *z*-score is separately computed for baseline and endline samples. For the GTEC score, we only have observations at the endline; *z*-scores are computed using the means and standard deviations among the endline samples. ***, **, and * indicate 1%, 5%, and 10% levels of statistical significance, respectively.
Source: Authors' calculations.

the experiment, the curriculum focuses on exam preparation even for first-year students (Sasaki 2018). Hence, panel A appears to suggest that our program helped mitigate the worsening attitudes among sampled students by stimulating their interest in an international vocation and international affairs (subcomponents 2 and 4, respectively).

The second motivational variable, WTC, also has significant and complex relationships with second-language learner confidence, motivation, and actual language use (e.g., MacIntyre 2007). As in the case of international posture, we took the eight items that measured WTC from the above-mentioned homepage because they have been successfully used in the past with Japanese high school students learning English as a second language (e.g., Yashima 2009).[11] The questions asked whether the students would be willing to communicate in English in hypothetical situations such as "group discussions on an English course," "giving a speech in public," and "a chance meeting with a foreign friend in the street." A six-point Likert scale offered the following choices: always, usually, sometimes, not very often, seldom, and never. We assigned 5 points to the answer always, 4 to usually, 3 to sometimes, 2 to not very often, 1 to seldom, and 0 to never, and computed the *z*-value of the total points.

The means of the *z*-scores are reported toward the bottom of panel A in Table 4. Similar to international posture, the control mean dropped from the baseline to the endline. However, the drop was smaller among the treatment group, and the initially nondifferent means became marginally different in the endline. This finding suggests that although the students' WTC tended to decline as a result of an English curriculum, such as the one followed in the top-tier high school under study, the Skype program played a role in mitigating the declining WTC.

As an additional variable to examine the attitudes of sample students, we use the Cambodia study tour dummy variable reported at the bottom of panel A. The school organized a 1-week study tour to Cambodia in December 2017 and the students had a chance to voluntarily apply for inclusion. The school provided us with a list of students who applied, and we constructed a dummy variable that

[11] Tomoko Yashima. WTC Scale. http://www2.ipcku.kansai-u.ac.jp/~yashima/data/wtc_scale.pdf (accessed April 15, 2019).

equals 1 if a student applied and 0 otherwise. Sixteen (10.1%) of the treated students and 11 (6.8%) of the control students applied. Although the difference is not statistically significant, the application rate was 4.2 percentage points higher among the treatment group. Importantly, the correlation between the application dummy and the total endline international posture score was positive with a correlation coefficient of 0.21 (not reported). Thus, the ICT program may have encouraged more students to apply by improving their international posture, which we may not be able to detect because of the weak statistical power.

B. Descriptive Analyses: English Communication Abilities

To quantify the students' English abilities, we use three sets of English tests: Versant, Benesse, and GTEC. We conducted the Versant tests both before and after our intervention to measure the development. In addition, the Benesse test was taken soon after our intervention started and toward the end of it, so the Benesse test score can also be used for the comparison using a DiD design. The GTEC test only measures cross-sectional differences after the intervention. All the test scores are presented as standardized *z*-scores. The scores of the standardized Versant test are comparable over time, and we computed *z*-values using the means and standard deviations among the baseline samples. Thus, we can measure the improvement in English communication abilities by looking at the changes in those abilities. However, the Benesse test score differs from one round to the other, as it is designed in accordance with the school curriculum and the difficulty of the test increases as students proceed with the curriculum. Thus, the *z*-scores are computed separately for the baseline and endline samples, and the changes in the *z*-scores before and after the treatment do not necessarily indicate changes in students' levels of English abilities because the Benesse test is likely to be more difficult in the endline.

Panel B in Table 4 shows the results of the treatment and control groups' respective scores in the international posture and English tests. Although we primarily intended to use the Versant test as our measure of English communication abilities, the answers provided by some students were not properly recorded because of overburdened internet connections. That is, the test was conducted in a computer room inside the school in order to provide the same test-taking environment for all students, but we ultimately organized a follow-up session for the students whose answers were not recorded. Because not all students attended the follow-up session, the problem is that scores were unrecorded for students who were less confident and more hesitant to retake the test. Appendix Table A2 presents the regression results, where the left-hand-side variable is a dummy variable equal to 1 if the student took the Versant test. The results show that the Versant take-up was not correlated with the observable characteristics at the baseline, but was correlated with the baseline Versant score at the endline (column 6). This suggests that poorly

performing students were less likely to have taken the endline Versant test, and we should therefore interpret the results cautiously.

For the Versant score, there is a slight difference between the two groups at the baseline, but it is not statistically significant. The score at the endline is statistically different between the two groups, with the treatment group having a higher score. However, this difference may be due to the types of students choosing to take the test, particularly among the treated students. Panel B also shows that the control mean increased from −0.093 to 0.406, which is a one-half standard deviation increase over 6 months. This is equivalent to a 2-point increase in the Versant score (out of a full score of 80), which is quite large according to the service provider. This improvement is most likely the consequence of the regular curriculum. By contrasting this result with our discussion above, we argue that while the regular school curriculum was unsuccessful in making the students' motivation to learn English more internationally oriented, it did improve their English communication abilities. The Skype program has the potential to sustain the students' intrinsic motivation and therefore supplement the regular curriculum.

The mean scores of the Benesse test, reported in the middle of panel B, were balanced at the baseline and there was no significant difference at the endline. One possible reason for this null result is that the Benesse test primarily measures reading abilities, whose improvement was not the main focus of the Skype program. The same logic applies to the overall GTEC score, which comprehensively measures four English-language skills. Yet, even when we look at the subcomponents of the GTEC, there was no statistical difference in subcomponent 2 (listening ability) or in subcomponent 4 (speaking ability). Taken together, the results shown in Panel B suggest that our intervention did not improve the English communication abilities of the treated students.

C. Econometric Specification

To rigorously analyze the impacts of the online program by controlling the baseline level of outcome variables or other characteristics, we applied two econometric specifications: analysis of covariance (ANCOVA) and DiD regression. Let y_{ijkt} be an outcome variable of student i in classroom j with English teacher k at time t. The ANCOVA specification is written as

$$y_{ijkt} = \alpha + \beta Treatment_j + \gamma y_{ijkt-1} + \eta_k + \varepsilon_{ijkt} \quad (1)$$

where $Treatment_j$ is a dummy variable equal to 1 for the student in treated class j, y_{ijkt-1} is an outcome variable at $t-1$ (since we have only two time periods, $t-1$ represents the baseline and t the endline), η_k is a set of English teacher dummies, and ε_{ijkt} is a heteroscedasticity-robust standard error. The standard error is not clustered because the number of clusters is much smaller than the rule-of-thumb number of

42 (Angrist and Pischke 2009). To control for possible intracluster correlations, together with correcting for the small number of clusters, we report the 95% confidence intervals (CIs) based on the wild cluster bootstrap method suggested in Cameron, Gelbach, and Miller (2008). We used *boottest* Stata command developed by Roodman et al. (2019) for the computation of the bootstrapped CIs.

In equation (1), β is the parameter of interest, which captures the intention-to-treat (ITT) impacts of the program. In addition to the ANCOVA specification, we also estimate a standard DiD model to control for unobserved, time-invariant, student-level heterogeneity, υ_i, using the following specification:

$$y_{ijkt} = \alpha + \beta Treatment_j * Endline_t + \delta Endline_t + \upsilon_i + \varepsilon_{ijkt} \quad (2)$$

where $Endline_t$ is a dummy variable equal to 1 if the data are collected in the endline (i.e., after the intervention). β in equation (2) is the parameter of interest, whereas δ measures the changes in the outcome variable from the baseline to the endline, which are mainly consequences of the regular school curriculum, as well as other changes that are common to all students.[12]

To analyze the different impacts of the online program by level of utilization, we use an instrumental approach to estimate the LATE (Imbens and Angrist 1994). Specifically, we replace $Treatment_j$ in equations (1) and (2) with $Lessons_i^k$, which equals 1 if student i took at least k lessons during the intervention period. We use $Treatment_j$ as an instrument for $Lessons^k{}_i$ to estimate the program impact for students in compliance by changing the threshold number of lessons. Since the assignment of treatment was randomized and the control students could not take any lessons, $Treatment_j$ works as a valid instrument. We, however, suffer from the weak instrument problem since the take-up rate was not high. To correct for this problem, we report the 95% CIs based on the wild cluster bootstrap because it also corrects for weak instruments (Roodman et al. 2019). In addition, we perform the conditional likelihood ratio tests developed by Moreira (2003), using *condivreg* Stata command by Moreira and Poi (2003) for robustness check.

D. Econometric Analyses: Intention to Treat

Table 5 shows the ITT estimates of the program impacts. Odd-numbered columns present the ANCOVA estimation results based on equation (1), while even-numbered columns present the DiD results based on equation (2). Panel A presents the estimated impacts on the attitude measures. Column 2 shows the positive and significant coefficients of the treatment on the total international posture score and the wild cluster bootstrap CI excludes 0, supporting our

[12] According to McKenzie (2012), ANCOVA analysis would be beneficial in power rather than DiD analysis when autocorrelations are low. The autocorrelation in our analysis ranged from 0.4 to 0.8, which is neither high nor low. We thus provide the results from both the ANCOVA and DiD analyses in Table 5.

Table 5. **Impacts of Online Program (Intention-to-Treat Estimation)**

A. Attitudes						
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
	Total International Posture		**Willingness to Communicate**		**Cambodia Tour (1 = yes)**	
	ANCOVA	**DiD**	**ANCOVA**	**DiD**	**OLS**	**OLS (with control)[a]**
Treatment (1 = yes)	0.15*		0.13		0.033	0.039
	(1.88)		(1.45)		(1.05)	(1.14)
Treatment × Endline (1 = yes)		0.12		0.078		
		(1.41)		(0.81)		
Baseline outcome	0.78***		0.64***			
	(22.61)		(13.96)			
Endline (1 = yes)		−0.11**		−0.23***		
		(−2.09)		(−3.16)		
English teacher B (1 = yes)	0.11		0.014		−0.028	−0.049
	(0.94)		(0.11)		(−0.59)	(−0.95)
English teacher C (1 = yes)	−0.13		−0.16		−0.053	−0.070
	(−1.24)		(−1.22)		(−1.17)	(−1.39)
English teacher D (1 = yes)	−0.21*		−0.14		−0.043	−0.071
	(−1.80)		(−1.13)		(−0.92)	(−1.40)
Wild cluster bootstrap (95% CI)	[0.09 0.20]	[−0.09 0.34]	[−0.05 0.32]	[−0.14 0.31]	[−0.01 0.08]	[0.00 0.08]
No. of observations	308	627	303	622	320	292

B. International posture (subcomponent)

	(1) Sub 1. Intercultural Orientation	(2)	(3) Sub 2. International Vocation	(4)	(5) Sub 3. Different Customs	(6)	(7) Sub 4. Foreign Affairs	(8)
	ANCOVA	DiD	ANCOVA	DiD	ANCOVA	DiD	ANCOVA	DiD
Treatment (1 = yes)	0.019		0.17**		0.020		0.25**	
	(0.21)		(2.21)		(0.19)		(2.56)	
Treatment × Endline (1 = yes)		−0.010		0.16*		−0.0087		0.18*
		(−0.11)		(1.87)		(−0.07)		(1.71)
Baseline outcome	0.70***		0.73***		0.41***		0.57***	
	(16.10)		(19.61)		(7.44)		(12.01)	
Endline (1 = yes)		−0.12*		−0.12**		−0.012		−0.0069
		(−1.82)		(−2.08)		(−0.13)		(−0.10)
English teacher B (1 = yes)	0.043		0.051		0.22		0.095	
	(0.35)		(0.47)		(1.49)		(0.69)	
English teacher C (1 = yes)	−0.11		−0.021		−0.077		−0.13	
	(−0.89)		(−0.19)		(−0.55)		(−1.00)	
English teacher D (1 = yes)	−0.15		−0.080		−0.18		−0.18	
	(−1.19)		(−0.68)		(−1.22)		(−1.31)	
Wild cluster bootstrap (95% CI)	[−0.31 0.34]	[−0.26 0.21]	[0.13 0.22]	[0.01 0.33]	[−0.18 0.21]	[−0.25 0.25]	[0.04 0.48]	[−0.01 0.39]
No. of observations	309	628	309	628	308	627	309	628

C. English communication abilities

	(1)	(2)	(3)	(4)	(5)	(6)
	(i) Total Versant		(ii) Benesse		(iii) Total GTEC	
	ANCOVA	DiD	ANCOVA	DiD	OLS	OLS (with control)[a]
Treatment (1 = yes)	0.099		−0.0051		0.0034	−0.0035
	(1.09)		(−0.06)		(0.03)	(−0.03)
Treatment × Endline (1 = yes)		0.042		0.018		
		(0.44)		(0.20)		
Baseline outcome	0.78***		0.71***			
	(17.38)		(16.72)			
Endline (1 = yes)		0.52***		−0.013		
		(8.41)		(−0.21)		
English teacher B (1 = yes)	−0.27**		0.0090		0.052	0.020
	(−2.13)		(0.08)		(0.31)	(0.12)
English teacher C (1 = yes)	−0.31**		−0.0056		0.0097	−0.081
	(−2.55)		(−0.05)		(0.06)	(−0.50)
English teacher D (1 = yes)	−0.44***		−0.24*		−0.19	−0.26
	(−3.37)		(−1.96)		(−1.16)	(−1.50)
Wild cluster bootstrap (95% CI)	[−0.09 0.32]	[−0.29 0.36]	[−0.22 0.19]	[−0.21 0.25]	[−0.12 0.12]	[−0.15 0.14]
No. of observations	243	553	312	627	318	291

D. GTEC test (subcomponent)

	(1) Sub 1. Reading	(2) Sub 1. Reading	(3) Sub 2. Listening	(4) Sub 2. Listening	(5) Sub 3. Writing	(6) Sub 3. Writing	(7) Sub 4. Speaking	(8) Sub 4. Speaking
	OLS	OLS (with control)[a]	OLS	OLS (with control)[a]	OLS	OLS (with control)[a]	OLS	OLS (with control)[a]
Treatment (1 = yes)	−0.023	−0.011	0.067	0.038	0.048	0.061	−0.021	−0.012
	(−0.20)	(−0.10)	(0.60)	(0.32)	(0.43)	(0.50)	(−0.19)	(−0.10)
English teacher B (1 = yes)	0.0040	−0.010	0.20	0.17	−0.10	−0.14	0.020	−0.0037
	(0.02)	(−0.06)	(1.19)	(0.95)	(−0.71)	(−0.87)	(0.12)	(−0.02)
English teacher C (1 = yes)	0.053	−0.049	0.026	−0.020	−0.15	−0.13	0.051	−0.053
	(0.35)	(−0.31)	(0.16)	(−0.12)	(−0.93)	(−0.74)	(0.33)	(−0.33)
English teacher D (1 = yes)	−0.20	−0.28*	−0.031	−0.077	−0.12	−0.11	−0.22	−0.29*
	(−1.28)	(−1.69)	(−0.19)	(−0.44)	(−0.78)	(−0.66)	(−1.34)	(−1.71)
Wild cluster bootstrap (95% CI)	[−0.19 0.17]	[−0.12 0.10]	[−0.11 0.27]	[−0.11 0.20]	[−0.48 0.57]	[−0.48 0.60]	[−0.18 0.18]	[−0.14 0.11]
No. of observations	265	243	553	265	243	553	265	243

ANCOVA = analysis of covariance, CI = confidence interval, DiD = difference-in-differences, GTEC = Global Test of English Communication, ITT = intention to treat, OLS = ordinary least squares.

Notes: Estimated coefficients are reported. ***, **, and * indicate 1%, 5%, and 10% levels of statistical significance, respectively. Numbers in parentheses are *t*-statistics based on heteroscedasticity-robust standard errors.

[a]In OLS (with control), the control variables in column 2 of Table 4 are added. Wild cluster bootstrap (95% CI) is for the treatment or the treatment × endline variable. Using *boottest* Stata command developed by Roodman et al. (2019), we implemented wild cluster bootstrapping with 1,000 replications. In so doing, we used the gamma distribution with the shape parameter of 4 and the scale parameter of 0.5 as weight for constructing the bootstrap samples.

Source: Authors' calculations.

discussion in the previous section. In the DiD estimation reported in column 2, the impact is positive but insignificant although the *t*-statistic is as large as 1.41, with the corresponding *p*-value of 0.148 (not reported). The point estimate is 0.12 and that of *Endline* is −0.11, which is statistically significant; these coefficients suggest that the overall international posture score declined from the baseline survey in June 2015 to the endline survey in December of the same year, but the Skype program offset the declining international posture score among the treated students. Furthermore, the significant teacher dummy suggests the presence of substantial teacher heterogeneity, as discussed in section III.B.

We report our results on WTC in columns 3 and 4. While not statistically significant, the point estimate is positive in both the ANCOVA and DiD estimations. In columns 5 and 6, we report results on the Cambodia tour. The point estimate is not significant, but the CI barely includes 0 in column 5 and excludes 0 in column 6. Hence, the treated students were more likely to have voluntarily applied for the opportunity to study abroad.

Panel B shows positive and significant impacts on subcomponents 2 (columns 3 and 4) and 4 (columns 7 and 8). The CIs for these two subcomponents exclude 0 (except for column 8, where the CI barely includes 0). With the point estimates for subcomponents 1 and 3 being close to 0, the impact on international posture comes from the changes in subcomponents 2 and 4. In particular, we find that while the Grade 10 students tended to become less interested in an international vocation—the size of the effect being 0.12 standard deviations (see column 4)—such a tendency was compensated for by our intervention.

Panel C of Table 5 shows the ITT estimates of the program impacts on students' English communication abilities in the same manner as panel A. The point estimates are small or even negative, particularly for the Benesse (columns 3 and 4) and GTEC tests (columns 5 and 6), and the corresponding *t*-statistics are close to 0. In addition, all the CIs include zero. Even if we look at the subcomponents of the GTEC shown in panel D, particularly subcomponents 2 (listening) and 4 (speaking), we find similar patterns of small coefficients with small *t*-statistics and CIs including zero. Hence, our regression analyses show that the Skype program had limited impacts on the students' English communication abilities.

However, attitudinal attributes have been reported to lead to eventual improvement in students' second-language skills (e.g., Sasaki 2011, Yashima 2002); therefore, the Skype program may have significant impacts over the long term. Unfortunately, all of the sample students had received the same amount of online intervention by the end of May 2016, and thus, we do not have variation to evaluate such long-term impacts. In addition, we may possibly have detected an effect if our intervention had been implemented for a longer period because Ross (2000), among others, finds that the duration is a major determinant of the effectiveness of second-language learning. Another important point to note from panel C is the significant coefficient of the endline dummy in column 2. As the scores of

the standardized Versant test are intertemporally comparable, the positive and significant coefficients suggest that students' communication abilities significantly improved over time, most likely due to the regular school curriculum in this top-tier high school.

E. Econometric Analyses: Local Average Treatment Effect

Table 6 reports the LATE estimates of program impacts on attitudes in panel A and on English communication abilities in panel B. In columns 1, 4, and 7 (where $k = 5$), the lesson dummy equals 1 if a student took at least five lessons in the intervention period; thus, the coefficient captures the impacts of the online program for students who completed at least five lessons.

In panel A, the size of the coefficient increases with k, indicating that the students who took more lessons benefited more from the program. For instance, the students who took 25 or more lessons (half of the recommended number by the service provider) have an international posture z-score that is 1.01 standard deviation higher than the average of the control students (column 3). However, the first-stage F-statistics decrease and the CIs widen as k increases because only 23 students (14%) completed 25 or more lessons, and the standard errors increase with k. This is one of the reasons why we do not find statistically significant coefficients for WTC (columns 4–6). In columns 7–9, although the coefficient is insignificant, CIs exclude or barely include zero, indicating the positive impact on students' participation in the overseas study.

In panel B, we find a similar increasing pattern for the Versant test (columns 1–3), but not for the Benesse test (columns 4–6) or the overall GTEC scores (columns 7–9). Unfortunately, none of the three indicators are a perfect measure of English communication abilities: (i) the Versant test with the nonrandom attrition, (ii) the Benesse test with the primary focus on reading skills, and (iii) the GTEC with the cross-sectional nature. Our tentative conclusion is that the impacts of our intervention on English communication abilities were at most limited.

F. Additional Analyses

We conducted two sets of additional analyses. First, we analyzed the heterogeneous treatment effects by interacting the treatment dummy with the control variables, including procrastination, gender, past exposure to English, family background, and baseline levels of the outcome variable. Panel A of Table 7 reports results for the international posture score; no interaction term is statistically significant, including those not reported (Table 7 only reports the results for the variables that were found to be correlated with some outcome variables in Appendix Table A1.) This may be because of the moderate size of the average treatment effects. Panel B reports the results for the Benesse test score. We found that

Table 6. **Impacts of Online Program (Local Average Treatment Effect Estimation)**

A. Attitudes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Total International Posture			Willingness to Communicate			Cambodia Tour (1 = yes)		
	IV k = 5	IV k = 10	IV k = 25	IV k = 5	IV k = 10	IV k = 25	IV k = 5	IV k = 10	IV k = 25
Lesson at least *k* times	0.29*	0.45*	1.01*	0.24	0.38	0.84	0.063	0.10	0.23
	(1.89)	(1.88)	(1.84)	(1.46)	(1.45)	(1.43)	(1.06)	(1.05)	(1.04)
Baseline outcome	0.77***	0.77***	0.79***	0.65***	0.65***	0.64***			
	(22.34)	(22.30)	(22.07)	(14.34)	(13.98)	(13.78)			
Teacher (strata) dummies	Y	Y	Y	Y	Y	Y	Y	Y	Y
First-stage *F*-statistics	37.2	15.1	5.1	40.7	15.5	5.4	47.5	18.9	6.7
Wild cluster bootstrap (95% CI)	[0.07 0.47]	[0.04 0.74]	[0.26 1.70]	[−0.03 0.56]	[−0.19 0.96]	[−0.07 1.95]	[0.01 0.10]	[−0.00 0.22]	[0.02 0.47]
Conditional LR test (95% CI)	[−0.01 0.60]	[−0.02 0.96]	[−0.05 2.34]	[−0.08 0.57]	[−0.14 0.92]	[−0.31 2.24]	[−0.06 0.18]	[−0.09 0.31]	[−0.21 0.72]
No. of observations	308	308	308	303	303	303	320	320	320

B. English communication abilities

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
		(i) Versant			(ii) Benesse			(iii) Total GTEC	
	IV k = 5	IV k = 10	IV k = 25	IV k = 5	IV k = 10	IV k = 25	IV k = 5	IV k = 10	IV k = 25
Lesson at least *k* times	0.18 (1.10)	0.32 (1.09)	0.79 (1.06)	−0.0096 (−0.06)	−0.016 (−0.06)	−0.034 (−0.06)	0.0066 (0.03)	0.011 (0.03)	0.023 (0.03)
Baseline outcome	0.77*** (16.61)	0.77*** (16.85)	0.78*** (17.56)	0.70*** (17.08)	0.70*** (17.06)	0.70*** (16.94)			
Teacher (strata) dummies	Y	Y	Y	Y	Y	Y	Y	Y	Y
First-stage *F*-statistics	31.2	10.8	3.1	37.3	14.4	5.4	48.7	19.0	6.7
Wild cluster bootstrap (95% CI)	[−0.20 0.50]	[−0.62 0.98]	[−0.92 2.35]	[−0.37 0.40]	[−0.64 0.60]	[−1.26 1.24]	[−0.23 0.39]	[−0.40 0.58]	[−0.85 1.41]
Conditional LR test (95% CI)	[−0.14 0.51]	[−0.24 0.93]	[−0.62 2.67]	[−0.32 0.29]	[−0.53 0.49]	[−1.19 1.10]	[−0.43 0.43]	[−0.70 0.70]	[−1.64 1.63]
No. of observations	243	243	243	312	312	312	318	318	318

CI = confidence interval, GTEC = Global Test of English Communication, IV = instrumental variable, LR = likelihood ratio.

Notes: Estimated coefficients are reported. ***, **, and * indicate 1%, 5%, and 10% levels of statistical significance, respectively. Numbers in parentheses are *t*-statistics based on heteroscedasticity-robust standard errors. In the IV estimations, the lesson dummy equals 1 if a student took at least *k* lessons, and this dummy variable is instrumented with the treatment dummy. Wild cluster bootstrap (95% CI) is for the lesson at least *k* times variable. Using *boottest* Stata command developed by Roodman et al. (2019), we implemented wild cluster bootstrapping with 1,000 replications. In so doing, we used the gamma distribution with the shape parameter of 4 and the scale parameter of 0.5 as weight for constructing the bootstrap samples. Conditional LR test is for the lesson at least *k* times variable, computed using *condivreg* Stata command developed by Moreira and Poi (2003).

Source: Authors' calculations.

Table 7. **Heterogeneous Treatment Effect**

A. International posture

X	(1) Procrastination	(2) Male (1 = yes)	(3) Been Abroad (1 = yes)	(4) Sports Club (1 = yes)	(5) Number of Books	(6) Baseline International Posture Score
Treatment (1 = yes)	0.14*	0.21*	0.24**	0.12	0.12	0.15*
	(1.83)	(1.95)	(2.41)	(0.95)	(0.77)	(1.87)
Treatment × X	−0.053	−0.12	−0.24	0.041	−0.0022	−0.11
	(−0.63)	(−0.72)	(−1.46)	(0.24)	(−0.04)	(−1.52)
X	−0.070	−0.053	0.32***	−0.029	0.047	N.A. (same as
	(−1.31)	(−0.50)	(2.75)	(−0.28)	(1.41)	baseline outcome)
Baseline outcome	0.78***	0.77***	0.75***	0.79***	0.80***	0.83***
	(22.53)	(21.73)	(19.62)	(22.37)	(23.50)	(17.58)
English teacher B	0.13	0.11	0.091	0.13	0.097	0.11
(1 = yes)	(1.15)	(0.97)	(0.81)	(1.11)	(0.88)	(0.94)
English teacher C	−0.13	−0.14	−0.12	−0.10	−0.092	−0.12
(1 = yes)	(−1.18)	(−1.24)	(−1.13)	(−0.97)	(−0.86)	(−1.15)
English teacher D	−0.20*	−0.21*	−0.22**	−0.19*	−0.18	−0.21*
(1 = yes)	(−1.78)	(−1.83)	(−1.98)	(−1.69)	(−1.55)	(−1.85)
No. of observations	303	308	308	303	301	308

B. Benesse score

X	(1) Procrastination	(2) Male (1 = yes)	(3) Been Abroad (1 = yes)	(4) Sports Club (1 = yes)	(5) Number of Books	(6) Baseline Benesse Score
Treatment (1 = yes)	−0.026	−0.055	−0.13	−0.054	−0.17	−0.0050
	(−0.32)	(−0.52)	(−1.29)	(−0.39)	(−1.12)	(−0.06)
Treatment × X	−0.10	0.099	0.31*	0.044	0.051	−0.019
	(−1.27)	(0.62)	(1.89)	(0.26)	(0.94)	(−0.23)
X	−0.0067	−0.068	−0.10	−0.016	0.0067	N.A. (same as
	(−0.11)	(−0.60)	(−0.82)	(−0.13)	(0.16)	baseline outcome)
Baseline outcome	0.69***	0.70***	0.70***	0.69***	0.68***	0.72***
	(15.75)	(16.47)	(16.62)	(15.74)	(15.18)	(11.82)
English teacher B	−0.033	0.0089	0.022	−0.021	−0.025	0.0095
(1 = yes)	(−0.28)	(0.08)	(0.18)	(−0.17)	(−0.20)	(0.08)
English teacher C	−0.067	−0.0045	0.0015	−0.036	−0.023	−0.0070
(1 = yes)	(−0.55)	(−0.04)	(0.01)	(−0.30)	(−0.20)	(−0.06)
English teacher D	−0.29**	−0.24*	−0.22*	−0.27**	−0.26**	−0.25**
(1 = yes)	(−2.26)	(−1.93)	(−1.74)	(−2.14)	(−2.08)	(−1.98)
No. of observations	306	312	309	305	303	312

X = control variables (i.e., procrastination, male, been abroad, sports club, number of books, and baseline Benesse score).
Notes: Estimated coefficients are reported. ***, **, and * indicate 1%, 5%, and 10% levels of statistical significance, respectively. Numbers in parentheses are *t*-statistics based on heteroscedasticity-robust standard errors.
Source: Authors' calculations.

only the interaction with the abroad dummy is positive and marginally significant, suggesting that the program may have widened the gap between strongly performing students with greater degrees of international exposure and those showing no such orientation because the former is more likely to take advantage of learning opportunities to further improve their English communication abilities.

The second set of analyses is the impact of the Skype program on the students' school performance based on their self-reported information. While admitting that we do not have more objective data based on assessments by their teachers, the treated students were more likely to work hard and actively participate in English classes at school (Table 8, columns 1–4). In addition, the treated students may be more likely to work hard in classes other than English classes (columns 5–6). Therefore, the program had positive impacts on overall school performance. In addition, the possibility of a crowding-out effect, where the students spend more time studying English while spending less time on other subjects, seems limited.

IV. Conclusion

We conducted a unique and rare field experiment in collaboration with a Japanese public high school to provide students with a home-use, ICT-assisted program for English. Through the examination of program usage records and panel data, we analyzed the factors associated with program utilization and estimated the program impacts. In our descriptive and econometric analyses, we found that the program significantly changed the internationally oriented attitudes of the treated students but not their English communication abilities. We could justifiably speculate that the insignificant improvement in their communication abilities was due to the low take-up rate of the targeted program. As we found that students showing a tendency to procrastinate were less likely to start and continue using the program, more research is warranted on how to improve and maintain students' motivation, particularly those with a tendency to procrastinate, and encourage them to use ICT-assisted programs such as the one targeted in this study. In addition, as improved internationally oriented attitudes could have a positive impact on students' English development on a long-term basis, future studies need to evaluate the long-term impacts of such programs.

We also found that although the entrance-exam-oriented regular school curriculum did improve the students' English (oral) communication abilities, it seemed to have negative effects on their international orientation. As we identified the positive causal effects of the online English learning program on the students' attitudes, given that it supplemented the weaknesses of the regular curriculum, future research should consider how to combine regular English lessons and such ICT-based programs in a complementary manner. In addition to encouraging interventions designed to encourage home use, using such programs during regular English lessons also might be an option.

Table 8. **Impacts on Self-Reported School Performance (Intention-to-Treat Estimation)**

	(1) I work hard in English classes	(2)	(3) I express my opinion in English classes	(4)	(5) I work hard in other classes	(6)	(7) I express my opinion in other classes	(8)
	OLS	OLS	OLS	OLS	OLS	OLS	OLS	OLS
Treatment (1 = yes)	0.29***	0.24***	0.22*	0.25**	0.27***	0.25***	0.025	0.13
	(3.10)	(2.97)	(1.93)	(2.56)	(2.73)	(2.69)	(0.22)	(1.27)
Baseline outcome		0.53***		0.53***		0.41***		0.48***
		(11.40)		(10.83)		(8.15)		(8.89)
English teacher B (1 = yes)	0.14	0.077	0.053	0.017	−0.013	0.016	0.17	0.15
	(1.02)	(0.69)	(0.30)	(0.12)	(−0.09)	(0.13)	(0.97)	(1.04)
English teacher C (1 = yes)	−0.051	−0.098	−0.14	−0.071	−0.13	−0.13	−0.077	0.0066
	(−0.38)	(−0.88)	(−0.78)	(−0.48)	(−0.93)	(−1.03)	(−0.44)	(0.05)
English teacher D (1 = yes)	−0.030	−0.056	−0.24	−0.28**	−0.094	−0.081	−0.077	−0.21
	(−0.21)	(−0.47)	(−1.44)	(−1.99)	(−0.66)	(−0.61)	(−0.45)	(−1.42)
Control mean at baseline	4.0		3.0		3.9		3.2	
Wild cluster bootstrap (95% CI)	[0.11 0.48]	[0.08 0.40]	[−0.00 0.44]	[0.06 0.44]	[0.07 0.48]	[0.07 0.43]	[−0.19 0.24]	[−0.07 0.34]
No. of observations	310	302	309	299	310	302	308	298

CI = confidence interval, OLS = ordinary least squares.

Notes: All outcomes are measured using a five-point Likert scale with 1 = not at all, 2 = no, 3 = neutral, 4 = yes, and 5 = definitely yes. Estimated coefficients reported. ***, **, and * indicate 1%, 5%, and 10% levels of statistical significance, respectively. Numbers in parentheses are *t*-statistics based on heteroscedasticity-robust standard errors. Wild cluster bootstrap (95% CI) is for the treatment variable. Using *boottest* Stata command developed by Roodman et al. (2019), we repeated wild cluster bootstrapping for 1,000 times. In so doing, we used the gamma distribution with the shape parameter of 4 and the scale parameter of 0.5 for weights.

Source: Authors' calculations.

References

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Bulman, George, and Robert W. Fairlie. 2016. "Technology and Education: Computers, Software, and the Internet." In *Handbook of the Economics of Education, Volume* 5, edited by Eric A. Hanushek, Stephen J. Machin, and Ludger Woessmann, 239–80. Amsterdam: Elsevier.

Cameron, Colin A., Jonah B. Gelbach, and Douglas L. Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *Review of Economics and Statistics* 90 (3): 414–27.

Dörnyei, Zoltán, and Stephen Ryan. 2015. *The Psychology of the Language Learner Revisited*. London: Routledge.

Duckworth, Angela L., Katherine L. Milkman, and David Laibson. 2018. "Beyond Willpower: Strategies for Reducing Failures of Self-Control." *Psychological Science in the Public Interest* 19 (3): 102–29.

Gee, James P., and Elisabeth R. Hayes. 2011. *Language and Learning in the Digital Age*. London: Routledge.

Glewwe, Paul, Michael Kremer, Sylvie Moulin, and Eric Zitzewitz. 2004. "Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya." *Journal of Development Economics* 74 (1): 251–68.

Golsteyn, Bart H. H., Hans Grönqvist, and Lena Lindahl. 2014. "Adolescent Time Preferences Predict Lifetime Outcomes." *Economic Journal* 124 (580): F739–F761.

Honda, Yuki, and Hiroshi Nishijima. 2007. "Survey of Lifestyles, Behaviors, and Attitudes of High-Schoolers in Tokyo." (Japanese). http://berd.benesse.jp/berd/center/open/report/toritsu_kousei/2009/pdf/siryou_01.pdf (accessed April 15, 2019).

Imbens, Guido M., and Angrist, Joshua D. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–76.

Imbens, Guido M., and Jeffrey M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47 (1): 5–86.

Kawaguchi, Daiji. 2016. "Fewer School Days, More Inequality." *Journal of the Japanese and International Economies* 39: 35–52.

Kim, Kimin, and Myoung-jae Lee. 2019. "Difference in Differences in Reverse." *Empirical Economics* 57 (3): 705–25.

Levy, Mike. 2009. "Technologies in Use for Second Language Learning." *Modern Language Journal* 93 (s1): 769–82.

MacIntyre, Peter D. 2007. "Willingness to Communicate in the Second Language: Understanding the Decision to Speak as a Volitional Process." *Modern Language Journal* 91 (4): 564–76.

McKenzie, David. 2012. "Beyond Baseline and Follow-up: The Case for More T in Experiments." *Journal of Development Economics* 99 (2): 210–21.

Ministry of Education, Culture, Sports, Science and Technology, Government of Japan (MEXT). 2015a. "Results of the English Test Conducted to Improve English Education in Japan in 2015." (Japanese). http://www.mext.go.jp/a_menu/kokusai/gaikokugo/1358258.htm (accessed April 15, 2019).

_____. 2015b. "Selection of 2015 Super-Global High Schools." (Japanese). http://www.mext.go.jp/a_menu/kokusai/sgh/1356366.htm (accessed April 21, 2019).

Moreira, Marcelo J. 2003. "A Conditional Likelihood Ratio Test for Structural Models." *Econometrica* 71 (4): 1027–48.

Moreira, Marcelo J., and Brian P. Poi. 2003. "Implementing Tests with Correct Size in the Simultaneous Equations Model." *Stata Journal* 3 (1): 57–70.

Onji, Kazuki, and Rina Kikuchi. 2011. "Procrastination, Prompts, and Preferences: Evidence from Daily Records of Self-Directed Learning Activities." *Journal of Socio-Economics* 40 (6): 929–41.

Ortega, Lourdes, and Gina Iberri-Shea. 2005. "Longitudinal Research in Second Language Acquisition: Recent Trends and Future Direction." *Annual Review of Applied Linguistics* 25: 26–45.

Osaka University. 2013. "Survey of Preference Parameters at Osaka University." (Japanese). http://www.iser.osakau.ac.jp/survey_data/doc/japan/questionnaire/japanese/2013QuestionnaireJAPAN.pdf (accessed April 15, 2019).

Pearson Inc. 2008. "Consistency of Versant English Test Scores Over Multiple Administrators." Unpublished.

Roodman, David, James G. MacKinnon, Morten Ørregaard Nielsen, and Matthew D. Webb. 2019. "Fast and Wild: Bootstrap Inference in Stata Using Boottest." *Stata Journal* 19 (1): 4–60.

Ross, Steven. 2000. "Individual Differences and Learning Outcomes on the Certificate of Spoken and Written English." In *Studies in Immigrant English Language Assessment*, edited by Geoff Brindley, 191–214. Sydney: NCELTR.

Sasaki, Miyuki. 2011. "Effects of Varying Lengths of Study-Abroad Experiences on Japanese EFL Students' L2 Writing Ability and Motivation: A Longitudinal Study." *TESOL Quarterly* 45 (1): 81–105.

_____. 2018. "Application of Diffusion of Innovation Theory to Educational Accountability: The Case of EFL Education in Japan." *Language Testing in Asia* 8 (1): 1–18.

Snilstveit, Birte, Jennifer Stevenson, Radhika Menon, Daniel Phillips, Emma Gallagher, Maisie Geleen, Hannah Jobse, Tanja Schmidt, and Emmanuel Jimenez. 2016. *The Impact of Education Programmes on Learning and School Participation in Low- and Middle-Income Countries: A Systematic Review Summary Report. 3ie Systematic Review Summary* 7. London: International Initiative for Impact Evaluation (3ie).

Yashima, Tomoko. 2002. "Willingness to Communicate in a Second Language: The Japanese EFL Context." *Modern Language Journal* 86 (1): 54–66.

_____. 2009. "International Posture and the Ideal L2 Self in the Japanese EFL Context." In *Motivation, Language Identity, and the L2 Self*, edited by Zoltán Dörnyei and Ema Ushioda, 144–63. Clevedon: Multilingual Matters.

_____. 2009. Kokusai. http://www2.ipcku.kansai-u.ac.jp/~yashima/data/kokusai.pdf (accessed April 15, 2019).

_____. 2009. WTC Scale. http://www2.ipcku.kansai-u.ac.jp/~yashima/data/wtc_scale.pdf (accessed April 15, 2019).

Yashima, Tomoko, Lori Zenuk-Nshide, and Kazuaki Shimizu. 2004. "The Influence of Attitudes and Affect on Willingness to Communicate and Second Language Communication." *Language Learning* 54 (1): 119–52.

Appendix

Table A1. **Baseline Correlates of Outcome Variables (Ordinary Least Squares Estimation)**

	(1) Total International Posture	(2) Willingness to Communicate	(3) Versant Score	(4) Benesse Score
Treatment	0.074	0.017	0.12	−0.10
(1 = yes)	(0.65)	(0.14)	(1.08)	(−0.88)
Procrastination	−0.096	−0.15***	−0.077	−0.067
[*z*-score]	(−1.58)	(−2.67)	(−1.17)	(−1.25)
Male	−0.22*	0.030	0.28	0.20
(1 = yes)	(−1.85)	(0.25)	(1.54)	(1.37)
English since Grade 3 or 4	0.051	−0.10	−0.013	−0.10
(1 = yes)	(0.41)	(−0.84)	(−0.09)	(−0.82)
English since Grade 5 or later	0.012	−0.19	−0.095	−0.23
(1 = yes)	(0.08)	(−1.11)	(−0.62)	(−1.45)
Been abroad	0.62***	0.43***	0.29**	0.0054
(1 = yes)	(5.50)	(3.62)	(2.11)	(0.04)
Own room	0.13	0.36**	0.15	−0.13
(1 = yes)	(0.65)	(2.25)	(0.97)	(−1.00)
Own personal computer	0.35*	0.094	0.62	0.34
(1 = yes)	(1.79)	(0.46)	(1.55)	(1.42)
Own tablet	0.10	0.19	0.085	0.16
(1 = yes)	(0.69)	(1.34)	(0.41)	(1.11)
Commuting time 21–40 minutes	−0.20	−0.092	−0.042	0.13
(1 = yes)	(−1.36)	(−0.66)	(−0.19)	(0.74)
Commuting time 41–60 minutes	−0.12	−0.058	−0.086	−0.15
(1 = yes)	(−0.69)	(−0.36)	(−0.42)	(−0.86)
Commuting time 61 minutes	0.24	0.26	0.11	−0.15
or over (1 = yes)	(1.21)	(1.32)	(0.51)	(−0.70)
Sports club	−0.0061	0.27**	0.10	−0.080
(1 = yes)	(−0.05)	(2.12)	(0.55)	(−0.54)
Number of books	0.021	0.054	0.081**	0.052
[1–6]	(0.56)	(1.32)	(2.26)	(1.28)
English teacher B	0.11	0.12	0.088	−0.069
(1 = yes)	(0.70)	(0.74)	(0.44)	(−0.39)
English teacher C	−0.041	0.11	0.072	−0.17
(1 = yes)	(−0.25)	(0.73)	(0.41)	(−0.90)
English teacher D	0.16	0.19	0.12	−0.21
(1 = yes)	(0.99)	(1.09)	(0.65)	(−1.25)
R-squared	0.149	0.141	0.122	0.078
Adjusted *R*-squared	0.096	0.087	0.061	0.020
No. of observations	291	292	262	289

Notes: Estimated coefficients are reported. ***, **, and * indicate 1%, 5%, and 10% levels of statistical significance, respectively. Numbers in parentheses are *t*-statistics based on heteroscedasticity-robust standard errors. The base category for the English-since variable is "English since Grade 1 or 2," for the commuting time variable it is "Commuting time 20 minutes or less," and for the teacher dummies it is "Teacher A."
Source: Authors' calculations.

Appendix A2. **Versant Take-Up (Ordinary Least Squares Estimation)**

	(1)	(2)	(3)	(4)	(5)	(6)
	= 1 if scored in Versant test					
	Baseline			Endline		
Treatment	−0.020	−0.012	−0.089***	−0.093**	−0.077**	−0.081**
(1 = yes)	(−0.60)	(−0.34)	(−2.65)	(−2.50)	(−2.21)	(−2.06)
Procrastination		0.0018		0.0044		0.0077
[*z*-score]		(0.09)		(0.27)		(0.45)
Male		0.052		−0.016		−0.012
(1 = yes)		(1.15)		(−0.39)		(−0.29)
English since Grade 3 or 4		−0.037		−0.0020		−0.018
(1 = yes)		(−0.99)		(−0.05)		(−0.45)
English since Grade 5 or later		−0.015		0.0020		−0.0050
(1 = yes)		(−0.26)		(0.04)		(−0.10)
Been abroad		0.023		−0.037		−0.047
(1 = yes)		(0.63)		(−0.95)		(−1.12)
Own room		0.042		−0.013		−0.024
(1 = yes)		(0.77)		(−0.28)		(−0.51)
Own personal computer		−0.095		0.0024		−0.0011
(1 = yes)		(−1.21)		(0.05)		(−0.02)
Own tablet		−0.079		0.063*		0.050
(1 = yes)		(−1.46)		(1.75)		(1.35)
Commuting time 21–40 minutes		0.036		0.067		0.067
(1 = yes)		(0.73)		(1.33)		(1.31)
Commuting time 41–60 minutes		0.087		0.014		0.022
(1 = yes)		(1.63)		(0.24)		(0.40)
Commuting time 61 minutes		0.048		0.059		0.054
or over (1 = yes)		(0.72)		(0.90)		(0.78)
Sports club		−0.064		−0.0098		−0.0069
(1 = yes)		(−1.37)		(−0.23)		(−0.16)
Number of books		−0.0033		0.015		0.020
[1–6]		(−0.24)		(1.11)		(1.46)
English teacher B	0.015	−0.0050	0.016	0.0077	−0.012	−0.028
(1 = yes)	(0.36)	(−0.11)	(0.37)	(0.17)	(−0.29)	(−0.64)
English teacher C	−0.010	−0.022	−0.022	−0.046	−0.042	−0.070
(1 = yes)	(−0.24)	(−0.49)	(−0.45)	(−0.90)	(−0.90)	(−1.47)
English teacher D	−0.081	−0.091*	−0.031	−0.041	−0.030	−0.039
(1 = yes)	(−1.60)	(−1.67)	(−0.64)	(−0.82)	(−0.62)	(−0.76)
Versant score in the baseline					0.030**	0.026
					(2.16)	(1.51)
R-squared	0.017	0.057	0.026	0.058	0.030	0.067
Adjusted *R*-squared	0.005	−0.001	0.013	−0.000	0.012	−0.002
No. of observations	320	292	320	292	288	262

Notes: Estimated coefficients reported. ***, **, and * indicate 1%, 5%, and 10% levels of statistical significance, respectively. Numbers in parentheses are *t*-statistics based on heteroscedasticity-robust standard errors. The base category for the English-since variable is "English since Grade 1 or 2," for the commuting time variable it is "Commuting time 20 minutes or less," and for the teacher dummies it is "Teacher A."
Source: Authors' calculations.