Experimental Economics (2021) 24:1047–1073
https://doi.org/10.1007/s10683-020-09689-1

ORIGINAL PAPER

Returns to effort: experimental evidence from an online language platform

Fulya Ersoy[1]

Received: 5 February 2020 / Revised: 6 November 2020 / Accepted: 8 November 2020 /
Published online: 21 November 2020

Abstract

While distance learning has become widespread, causal estimates regarding returns to effort in technology-assisted learning environments are scarce due to high attrition rates and endogeneity of effort. In this paper, I manipulate effort by randomly assigning students different numbers of lessons in a popular online language learning platform. Using administrative data from the platform and the instrumental variables strategy, I find that completing 9 Duolingo lessons, which corresponds to approximately 60 minutes of studying, leads to a 0.057–0.095 standard deviation increase in test scores. Comparisons to the literature and back-of-the-envelope calculations suggest that distance learning can be as effective as in-person learning for college students for an introductory language course.

Keywords Returns to effort · Distance learning · Manipulation of effort · Field experiment

JEL Classification I23 · I26 · C93

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10683-020-09689-1) contains supplementary material, which is available to authorized users.

✉ Fulya Ersoy
fulya.ersoy@lmu.edu

1 Department of Economics, Loyola Marymount University, 1 LMU Drive, Los Angeles, CA 90045, USA

1 Introduction

There has been a growing interest in the use of educational technology (Escueta et al. forthcoming) and a steady increase in the number of students who engage in distance learning in recent years (Seaman et al. 2018).[1] Although technology-assisted learning has the potential to be a transformative force in education due to its flexibility, accessibility (Goodman et al. 2019), and affordability (Deming et al. 2015), we do not know much about the effectiveness of effort in these learning environments. High attrition rates in distance learning environments[2] combined with endogenous effort choices of students render measuring the causal returns to effort in these settings quite difficult. In this paper, I design an experiment in which study effort is exogenously manipulated and cleanly observed to measure the returns to effort for one of the most popular language learning platforms, Duolingo. The aim of this paper is not to evaluate the effectiveness of Duolingo, it is rather to understand whether and how study effort affects performance in a remote setting with technology-assisted learning.

To measure causal returns to effort in a technology-assisted learning environment, I design an experiment. First, I recruit college students who want to learn Spanish. I assess their initial Spanish knowledge with two tests. Students take an *internal* test which is based on Duolingo lessons and an *external* test, WebCAPE. Then, students sign up for Duolingo. I randomly assign them to one of the five groups (*32-lessons*, *48-lessons*, *64-lessons*, *80-lessons*, and *96-lessons*) in which they are asked to complete a varying number of Duolingo lessons for four weeks. To increase compliance with assignments, I require students to commit to complete the number of lessons assigned, I send them email reminders twice a week about their progress, and I tie their compensation to their compliance with the lesson assignments. I observe students' effort within Duolingo through administrative data. After four weeks, students once again take the *internal* and *external* tests which assess their final Spanish knowledge. Both of these tests are relevant metrics of performance since the *internal* test is similar to performance measures in classroom settings (i.e. lecture-based assessment) and the *external* test is used by many post-secondary institutions to place students into the appropriate levels of language courses.

The experimental design is successful at manipulating Duolingo effort without causing differential attrition. Random assignment to different lesson groups generates the necessary exogenous variation in Duolingo effort. Most of the students complete their assigned number of lessons. The difference between the distributions of the number of completed lessons for any two assignment groups is statistically significant. Furthermore, there is no evidence of differential attrition across different assignment groups.

[1] For example, searches for learning related apps have grown 80% and searches for apps related to learning a language have grown 85% from 2016 to 2017 according to Google data.

[2] For example, Jordan (2015) reports that the average completion rates for Massive Online Open Courses is approximately 15% based on 217 courses from 15 different platforms.

I find that effort within Duolingo positively impacts test performance. It is often debated what test scores measure. Test scores are likely to be influenced by different factors such as innate ability, effort invested in studying, effort on the test, and random factors unrelated to the test. In this paper, I am able to show that effort invested in studying improves test scores. In particular, I find that completing 9 lessons, which corresponds to approximately 60 minutes of Duolingo, results in a 0.095 standard deviation (sd) increase in *internal* test scores (statistically significant at 1% level) and a 0.057 sd increase in *external* test scores (statistically significant at 5% level) using an instrumental variables approach.[3] Furthermore, I find that the effectiveness of Duolingo positively correlates with college GPA. The estimated effects are for intrinsically motivated college students for an introductory level course. Hence, the generalizability of the results to other populations or contexts should be explored.

The results suggest that effort in Duolingo can be as effective as effort in in-person courses. In particular, I find that a one sd increase in the number of lessons completed leads to a 0.19–0.32 sd increase in test scores. These effects are comparable to the effects reported in the literature for in-person (non-language) courses [0.20 sd in Eren and Henderson (2011), 0.25 sd in Bonesrønning and Opstad (2012), and 0.14 sd in Bonesrønning and Opstad (2015)]. Moreover, using WebCAPE's benchmark scores for placement into college-level language courses, a back-of-the-envelope calculation suggests that the effectiveness of Duolingo effort is comparable to the effectiveness of effort in in-person language courses. In particular, an individual needs to complete, on average, 38 hours with Duolingo to be placed in the second semester of college level Spanish course which is comparable to the time spent in a semester of in-person Spanish course. However, we should take this conclusion as suggestive due to the linear extrapolation assumption used for the back-of-the envelope calculations and the noisiness of the results for WebCAPE test scores.

This paper contributes to three areas in the literature. First, it contributes to the literature on returns to effort. A limited number of studies explore the causal relationship between study effort and educational outcomes due to the fact that effort is multi-dimensional, difficult to observe, and hard to manipulate in most settings. To overcome the difficulty of unobservability, researchers frequently use students' self-reports of effort (Krohn and O'Connor 2005; De Fraja 2010; Kuehn and Landeras 2012). However, self-reports can contain substantive reporting error (Stinebrickner and Stinebrickner 2004) and can suffer from social desirability bias (Nederhof 1985; Furnham 1986). In this study, I measure effort using administrative data on the number of lessons completed, which provides a quantitative measure of effort while controlling for quality due to the structure of Duolingo. To deal with the endogeneity of effort, researchers use potentially exogenous variation in the value of leisure (Stinebrickner and Stinebrickner 2008; Metcalfe et al. 2019), change financial

[3] Most researchers assume linear returns when modeling effort choices, but whether returns to effort are non-linear is not explored empirically. Although the experimental design of this paper is suitable for such an exploration, I am unable to infer conclusions about the non-linearity of returns to effort given the small sample size in each assignment group.

incentives associated with effort (Angrist and Lavy 2009; Fryer 2011; Hirshleifer 2016), modify non-financial incentives associated with effort (Bonesrønning and Opstad 2015; Chevalier 2018), control for student and teacher fixed effects (Eren and Henderson 2011), use natural variation in the amount of homework assigned or manipulate whether homework submission is compulsory (Betts 1996; Eren and Henderson 2008; Grodner and Rupp 2013), modify course materials covered in lectures (Chen and Lin 2008), and vary whether attendance is mandatory (Dobkin et al. 2010). In this study, I create exogenous variation in effort by randomly assigning different numbers of lessons to students in a remote setting. To my knowledge, this is the first study that estimates the causal returns to effort for a remote setting with technology-assisted learning.

Second, this paper contributes to the small but growing literature on technology-assisted learning. The literature on technology-aided instruction has mixed findings with widely varying impacts, from significantly negative effects to significantly positive effects (Glewwe and Muralidharan 2016). The highly effective interventions seem to be the ones that uses technology to personalize the learning to the level of the student (Banerjee et al. 2007; Barrow et al. 2009; Muralidharan et al. 2019) and the ineffective interventions are the ones that use technology as a substitute rather than as a complement to traditional instruction methods (Lai et al. 2015; Linden 2008). Most of this literature focuses on using technology in a school or after school context where the teachers and peers are available. Duolingo utilizes technology-aided instruction methods effectively as a remote learning platform. In particular, it personalizes and tailors the learning experience of its users through technology, it provides real time feedback, and it uses gamification features to improve attention. I contribute to this literature by showing that well-designed technology aided *online* instruction can be as effective as in-person learning as long as students exert effort.

Third, this paper contributes to the distance learning literature that explores the causes and remedies of low student engagement and high attrition rates (see Bawa 2016, for a literature review). In a survey of 1,698 learners in 20 different online courses, Kizilcec et al. (2015) identify time management as the main reason for attrition. As a remedy to these problems, researchers try using small nudges in online settings to improve course outcomes (Kizilcec and Brooks 2017).[4] My paper contributes to this literature by showing that lesson assignments coupled with modest monetary incentives and reminders increase student effort successfully in a remote environment.[5]

[4] For example, Renz et al. (2016) find that reminder emails about unseen lecture videos increase lecture views.

[5] This finding also relates to Clark et al. (2017). The authors show assigning task-based goals has large impacts on course performance of students in in-person classrooms.

2 Experimental design

The platform used in this experiment, Duolingo, is one of the most popular language learning platforms with 300 million users worldwide. In Duolingo, users learn by practice. Each lesson is a compilation of speaking, listening, translation, and multiple-choice questions. Users can make many mistakes within a lesson but need to eventually answer all of the questions correctly in order to complete the lesson. Duolingo has 317 Spanish lessons categorized into Beginner, Intermediate, and Advanced Skills.[6] Duolingo reports that a lesson typically takes 5 to 10 minutes to complete. Beginner lessons take less time than Advanced lessons to complete. Users need to complete all the previous lessons to be able to unlock the upcoming lessons.[7]

In this experiment, students from various public universities in California sign up for a language learning study and take an eligibility survey. Eligible participants receive the study instructions and complete a survey about their demographics, language background, schooling background, and personality traits. Then, they take tests which assess their initial Spanish knowledge and sign up for Duolingo. Participants are randomly assigned to one of the five groups. I ask each group to complete a different number of Duolingo lessons (ranges from 32 to 96 lessons in a month). After studying through Duolingo for four weeks, participants take tests that assess their final Spanish knowledge. They also complete a short survey which collects information about how and how much they study Spanish during the study. Participants who successfully complete the study receive a completion payment of $50, paid as Amazon e-gift cards. Figure 1 summarizes the experiment.

3 Recruitment and randomization

The experiment took place in January and February of 2017. I recruited intrinsically motivated students from large public universities in California.[8] To recruit students from these universities, I emailed a recruitment flyer to the faculty and staff

[6] Although the skill levels are named as beginner, intermediate, and advanced, Duolingo courses generally teach at most to the B1 level of Common European Framework of Reference for Languages (CEFR) which is threshold level independent user. For more details on CEFR standards, see https://rm.coe.int/1680459f97.

[7] See Appendix A.1 in ESM for the structure of Duolingo and Appendix A.2 in ESM for a list of Duolingo lessons.

[8] The recruitments took place in December 2016. The universities which I recruit from are California Polytechnic State University-San Luis Obispo, California State Polytechnic University-Pomona, California State University-Fresno, California State University-Fullerton, California State University-Los Angeles, California State University-Northridge, California State University-Sacramento, University of California-Davis, University of California-Irvine, University of California-Santa Barbara, University of California-San Diego, San Diego State University, and San Francisco State University. I choose a large number of universities since I need a large subject pool to use in both this paper and a companion paper. I determined this list using the College Navigator website based on the following criteria: being public, offering at least a Bachelor's degree, having at least 20,000 undergraduates and admitting 30% to 70% of its applicants, and being in California.

in all departments. The flyer specifically stated that the students who want to participate in the language learning study should be motivated to learn Spanish and should know no or little Spanish to start with (see Appendix A.3 in ESM for the recruitment flyer). Four hundred seventy four subjects signed up for the study and 305 subjects took the eligibility survey. The eligibility criteria are being a student in one of the universities from which I recruited, being age 18 or above, being highly interested in learning Spanish, being able to commit up to 4 hours to study Spanish online, and knowing no or little Spanish. Sixty two subjects failed at least one of the five eligibility criteria and were excluded from participating in the study.[9] Of the remaining 247, 216 subjects completed the initial survey, 158 subjects took the initial Spanish tests and 146 subjects joined Duolingo. These 146 subjects form the baseline sample.

To be able to measure the effect of effort on performance, an exogenous variation in effort is necessary. To create this variation, I randomly assign the participants to one of the five different lesson groups at the beginning of the study.[10] The participants in the *32-lessons* group are asked to complete 8 lessons per week for 4 weeks. Similarly, the participants in the *48-lessons*, *64-lessons*, *80-lessons* and *96-lessons* groups are asked to complete 12, 16, 20, 24 lessons per week for 4 weeks, respectively.[11] Participants do not know the number of lessons assigned to other groups but know that the group assignments are random. Since lesson assignments are on a weekly basis, the participants need to log into the platform at least 4 times during the study if they want to complete their assignments.[12]

To increase the compliance of the subjects with the assignments, I use complementary approaches. First, at the beginning of the study, I inform participants that it is crucial that they study as many lessons as they are assigned and complete their assignments on time for the results of the research study to be valid. Then, I ask them to type "I will complete my assigned lessons, no more, no less. I understand this is necessary for the validity of the results of this research study.".[13] This is similar to the "Implementation Intention" method introduced by Gollwitzer (1999) and

[9] The answers to the eligibility questions are self-reported. 12 subjects reported that they are not students in one of the universities listed in Footnote 8, 5 subjects reported that they are not highly interested in learning Spanish, 20 subjects reported that they are not able to commit up to 4 hours and 32 subjects reported that they know Spanish at the intermediate level or above.

[10] Once participants signed up for Duolingo, they were encouraged to take a very short Spanish test within Duolingo. Duolingo uses this test to suggest a starting level for its users and marks all the lessons up to that level as complete. I ask participants not to complete any lessons until they receive their lesson assignments and assign their lessons based on the results of this placement test. Hence, the assigned lessons differ within groups depending on the initial levels of the participants. Once the deadline for joining Duolingo passed, randomization to different groups was done using the list randomizer on random.org.

[11] Vesselinov and Grego (2012) measure how time spent in Duolingo correlates with improvement in test scores using a random representative sample of Duolingo users studying Spanish. According to their data, the mean hours of study for four weeks correspond to approximately 5 hours. I have determined the number of lessons assigned such that the time spent for *64-lessons* is similar to this mean.

[12] I choose not to intervene with the number of logins per week since such an intervention can create differential attrition across groups.

[13] The copy/paste function is disabled for this text and the participants need to type this sentence exactly as shown to be able to proceed to the next parts of the survey.

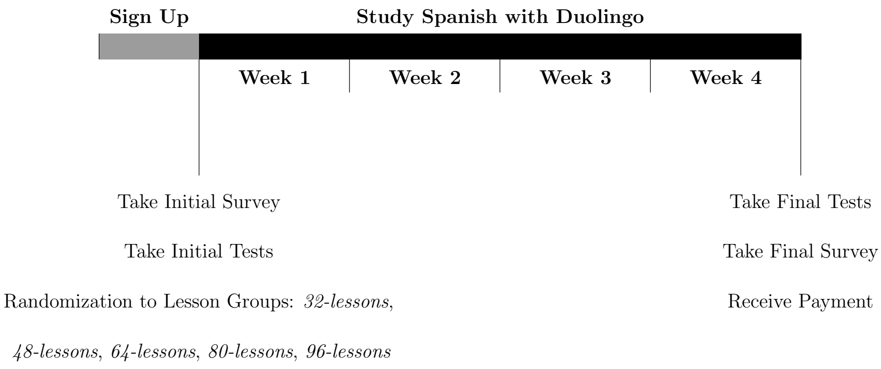

Fig. 1 Timeline of the experiment

known to be highly effective at influencing behavior (Kamenica 2012). Second, I send real-time warning e-mails to participants who complete more lessons than assigned and progress e-mails to all participants twice a week as a reminder of how many more lessons they need to complete.[14] Third, I tie compensation of the participants to their compliance with lesson assignments. From the beginning of the study, participants know that their compensation can be reduced depending on the severity of their non-compliance; but they are not told the details of the procedure.[15]

4 Data

4.1 Effort measures

The concept of study effort is not well defined and might mean different things in different contexts.[16] In self-paced learning environments like Duolingo, effort can be defined as the amount of work one puts into learning a subject (e.g. Spanish) in a given time. I have two effort measures: the number of lessons completed and time spent in Duolingo during the four-week period. For each participant, I observe the names of the lessons completed, whether those lessons are completed for the first time or repeated for practice, and timestamps. To create my main effort measure, I add up all the lessons participants completed during the four-week period.[17] As

[14] See Appendix A.4 and A.5 in ESM for screenshots of example e-mails.

[15] See Appendix A.6 in ESM for the exact wording of the instructions regarding the payment procedure. Following the instructions, I ask subjects to answer some comprehension questions about the payment procedure to check their understanding of the procedure. Subjects need to answer them correctly to continue. Appendix A.6 in ESM displays the screenshots of these questions.

[16] For example, Owen (1995) defines student effort as days of school attended per year and hours spent in the classroom or doing homework.

[17] This measure includes new as well as repeated lessons. Sixty nine percent of the subjects did not complete any repeated lessons and an additional 20 percent of them completed between one and five repeated lessons. Using the number of new lessons completed instead of the number of all lessons completed does not affect the results.

an alternative effort measure, I create minutes spent in Duolingo during the four-week period by using the timestamps.[18] Timestamps only indicate when a lesson was completed but do not provide information about when a lesson was started.

To construct the time spent variable for each lesson, I proxy the start time of a lesson with the end time of the previous lesson and calculate the difference between the end time of a lesson and the proxied start time. This exercise is justified given that most of the lesson assignments were completed in one sitting.[19] If the difference between start times of two successive lessons is more than 20 minutes or if the time spent variable is missing (which is the case for the very first lesson a participant completed), I replace the time spent with the median time spent per lesson, 3.97 minutes. Due to the limitations of time spent data and the advantages of the number of lessons completed data (being a quantitative cleanly observed measure), I use the number of lessons completed in Duolingo as my preferred effort measure.

4.2 Spanish assessment tests

I assess participants' Spanish knowledge with online language tests (internal and external) at the beginning and at the end of the study. The *internal* test allows me to test whether Duolingo is effective at teaching the content it aims to teach whereas the *external* test additionally allows me to analyze whether the contents of Duolingo are relevant to learn Spanish. I choose not to incentivize the participants based on their test scores since incentivizing the performance could decrease compliance with lesson assignments, induce substantial cheating in the test, and lead to differential attrition across ability types. Additionally, learners are not paid for performance in conventional environments and paying them might affect their intrinsic motivation.

I construct the *internal* Spanish test based on Duolingo lessons. This method is similar to the ones used in classrooms to measure performance. To construct the test, I scrape all questions from all Beginner and Intermediate Duolingo lessons. I create two separate tests based on these questions, Test A and Test B. A random half of participants receive Test A as the initial test and Test B as the final test and the order is switched for the other half.[20] Each test is checked for correctness by at least 3 native Spanish speakers.[21] Each test contains 80 questions (40 from Beginner Skills and 40 from Intermediate Skills) and the majority of the questions are multiple-choice questions with one or more correct answers. The scoring is done such that a participant

[18] I also have data on self-reported minutes of study on Duolingo for the last week of the experiment. Multiplying these self-reports by four and using them as the time spent measure instead of the administrative one leads to similar estimates.

[19] Appendix Figure B.1 in ESM displays a histogram of minutes spent per lesson based on timestamp data. Minutes spent per lesson are 20 or less in 82.4% of the lessons and 100 or more in 14.2% of the lessons.

[20] The tests are accessible through these links: http://mylmu.co1.qualtrics.com/jfe/form/SV_5p9uDIngEXud6LP (Test A) and http://mylmu.co1.qualtrics.com/jfe/form/SV_cOYMUgRrZMEts8d (Test B).

[21] I thank Jaime Arellano-Bover, Nano Barahona, Eduardo Laguna-Muggenburg, Jose Maria Barrero, Alejandro Martinez- Marquina, Oriol Pons-Benaiges, and Diego Torres-Patino for their help.

who answers all questions correctly would get a score of 1000 and a participant who answers all questions randomly would get a score of 0 on average.

I use WebCAPE (Web-Based Computer Adaptive Placement Exam) as an external test to measure Spanish knowledge. As the name suggests, WebCAPE is an adaptive test. Adaptive tests are common forms of language evaluation (e.g. TOEFL) and they are becoming more widespread in high school evaluations as more states use computer adaptive tests to evaluate core courses. WebCAPE is a placement test that is used by more than 200 post-secondary institutions, including some of the institutions in my sample. WebCAPE provides benchmark test scores that can be used to place students in different semesters of college level Spanish courses, which is one of the relevant margins for language learners. The test takes approximately 20 minutes. It has a high validity correlation coefficient with other Spanish tests and high test-retest reliability.

4.3 Summary statistics and balance

Appendix Table B.1 in ESM Column (1) shows the summary statistics of the sample who completed the initial survey and Column (2) shows the summary statistics of the baseline sample who joined Duolingo. Sixty-nine percent of the baseline sample is female, 36 percent is Caucasian, followed by 33 percent Asian. Average years of education for mother and father are 15 years and the average annual household income is around $103,000. Fifty-four percent of the participants major in Health and STEM, 29 percent of them major in Arts, Humanities and Social Sciences, and 10 percent of them major in Business and Economics. Comparing the initial sample (Column 1) to the baseline sample (Column 2), we see that Asians, students with higher SAT/ACT math scores, and students with a higher agreeableness score are more likely to join Duolingo (p-values are 0.029, 0.043, and 0.094, respectively as shown in Column 8) conditional on completing the initial survey.

To test balance across treatment groups in the baseline sample, I check whether observable characteristics are the same across groups. Appendix Table B.1 in ESM Columns (3)–(7) report the averages and Column (9) reports the p-value of the joint test of equality. One out of 30 variables (being Hispanic) differs across treatments at the 10% level, which is lower than what is expected by chance. Growth Mindset Score is also not perfectly balanced (p-value=0.133). According to the last two rows of Table B.1, mean initial scores for the internal and the external tests are not statistically different across treatment groups. To explore the balance of initial test scores across treatment groups further, Appendix Figure B.2a in ESM depicts the cumulative distribution of initial scores for the internal test (on the left) and for the external test (on the right). For the internal test scores, the distributions look quite similar to each other. For the external test scores, there is some suggestive evidence that the participants in the *80-lessons* group were worse than the others to start with.

4.4 Attrition

From the baseline sample of 146 subjects, 120 of them took both Spanish tests at the end of the study (end sample). Hence, the attrition rate is moderate (17.8%).[22] If the final Spanish test take-up differs across treatment groups, we will have differential attrition.[23] To check this possibility, I first look at the attrition rates across assignment groups. In the *32-lessons* group, 12 percent of the subjects did not take the final Spanish tests. The attrition rates are 15, 22, 19, and 20 percent for the *48-lessons*, *64-lessons*, *80-lessons*, and *96-lessons* groups, respectively. I cannot reject the joint equality of these attrition rates across groups (two-sided p-value from the linear hypothesis testing is 0.867). Nevertheless, I conduct a placebo check using the initial test scores as the dependent variable in the main estimating equation, present bounds for the treatment effects, and calculate inverse probability weighted estimates of the treatment effects as a robustness check in Sect. 5.5.1.

To further test for differential attrition, I check whether observable characteristics are the same across groups in the end sample. Table 1 Column (1) reports the averages for the end sample, Columns (2)–(6) report the averages for each treatment group, and Column (7) reports the p-value of the joint test of equality across treatment groups. Three out of 30 variables (being Hispanic, SAT/ACT reading scores, and growth mindset score) differ across treatments. The lack of balance for being Hispanic and growth mindset score in the end sample is not surprising given that these variables are not nicely balanced in the baseline sample. For the SAT/ACT reading scores, there seems to be a suggestive pattern such that students with lower SAT/ACT reading scores are less likely to take the final tests if they are in treatment groups with higher number of assignments. According to the last two rows of Table 1, mean initial scores for the internal and external tests are not statistically different across treatment groups. Nevertheless, to account for potential differences across groups in the end sample, I control for all of these observable characteristics in the regressions.

5 Findings

In this section, I first show that random assignment to different lesson groups is highly successful at manipulating effort levels of the participants (number of lessons completed and minutes spent) within Duolingo. Then, I present the effect of Duolingo effort on the internal and external test scores. Finally, I discuss heterogeneity of the effects, validity of the exclusion restriction, and robustness of the results.

[22] The overall attrition rate in this paper is similar to the 15% average attrition rate reported for field experiments published in top economics journals (Ghanem et al. 2019).

[23] Since I know the exact number of lessons subjects complete, subjects not completing any lessons or subjects stopping to complete lessons (differentially across assignment groups) does not cause a problem for the estimations as long as these subjects take the final Spanish tests.

Table 1 Summary statistics for end sample

	Averages						P-Values
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	End sample	*32-lessons* group	*48-lessons* group	*64-lessons* group	*80-lessons* group	*96-lessons* group	Joint test across groups
Female	.68	.61	.64	.67	.77	.68	.804
African-American	.08	.04	.09	.05	.12	.11	.85
Asian	.36	.26	.45	.57	.31	.25	.103
Caucasian	.36	.35	.41	.24	.35	.43	.71
Hispanic	.13	.26	0	.05	.08	.21	.028
Age	22.75	22	23	23.71	22.19	22.96	.707
Art, Humanities, and Soc. Sci.	.27	.13	.27	.33	.23	.36	.414
Business and Economics	.1	.13	.09	.1	.08	.11	.981
Health and STEM	.55	.57	.59	.52	.62	.46	.832
Freshman	.09	.13	.05	.1	.12	.07	.868
Sophomore	.12	.09	.09	.24	.08	.11	.448
Junior	.23	.26	.18	.38	.23	.14	.38
Senior	.38	.43	.5	.14	.38	.43	.143
GPA	3.43	3.33	3.35	3.38	3.49	3.57	.185
SAT/ACT math	.06	− .03	.1	.21	− .02	.09	.473
SAT/ACT reading	.07	− .12	.06	.16	− .01	.24	.098
SAT/ACT missing	.53	.35	.73	.48	.54	.57	.14
Mother's Education	14.6	13.65	14.18	15.52	14.62	15	.551
Father's Education	14.72	14.17	15.64	15.71	14.15	14.21	.277
Locus of Control	4.24	4.52	4.14	4.19	4.27	4.11	.939
Self Control	45.42	46	46.05	45.86	44.88	44.61	.961
Growth Mindset	28.94	27.57	29.27	28.52	30.08	29.07	.086

Table 1 (continued)

	Averages						P-Values
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	End sample	*32-lessons* group	*48-lessons* group	*64-lessons* group	*80-lessons* group	*96-lessons* group	Joint test across groups
Openness	7.27	6.87	7.27	7.33	7.65	7.18	.647
Neuroticism	5.85	5.35	6.05	5.43	6.19	6.11	.369
Conscientiousness	7.64	7.65	7.82	7.43	7.62	7.68	.934
Agreeableness	7.28	7.39	7.41	7.52	7.31	6.89	.644
Extroversion	5.88	5.87	5.45	6	6.42	5.61	.405
Annual Income	104,333	109,782	1,203,40	112,857	84,423	99,375	.865
Monthly Expenditure	236.67	250	251.14	215.48	209.62	255.36	.892
Target Spanish in a Month	169.98	150.39	164.82	175.05	171.96	184.46	.593
Target Spanish in a Year	272.32	265.87	267.18	264.24	277.62	282.79	.905
Initial Test Score (Internal)	325.65	320.26	321.42	298.86	297.69	379.46	.459
Initial Test Score (External)	152.78	132.78	160.14	155.76	120.35	191.32	.225
Observations	120	23	22	21	26	28	120

Omitted ethnicity and major categories are other, and omitted cohort category is masters or above. SAT/ACT math (reading) is equal to standardized SAT math (reading) score if the student reports SAT math (reading) score only, standardized ACT math (reading) score if the student reports ACT math (reading) score only, the average of standardized SAT and ACT math (reading) scores if the student reports both, and the group average of standardized SAT/ACT math (reading) if the student does not report any scores. SAT/ACT missing is equal to 1 if the student does not report any SAT/ACT scores. Mother's Education and Father's Education variables are based on subjects' answers to the question about their parents' highest level of education. "Locus of Control" is based on the Locus of Control Survey (Rotter (1966)) and ranges between 0 and 8. A higher number signifies an external locus of control. "Self Control" is based on the Self Control Survey (Tangney et al. (2004)) and ranges between 13 and 65. "Growth Mindset" is based on the Growth Mindset Survey (Paunesku et al. (2015)) and a higher score signifies growth mindset (range: 8 to 48). Extroversion, Agreeableness, Neuroticism, Conscientious and Openness are based on a short version of the Big Five Personality Trait Survey (Rammstedt and John (2007)) and the score ranges between 2 and 10 for each of them. Annual Income is based on subjects' answers to the question about their annual family income and Monthly Expenditure is based on subjects' answers to the question of how much money they spend on food, clothing, leisure each month. Target Spanish in a Month and in a Year show the level of Spanish subjects want to reach in a month and in a year, respectively (range: 0 to 400). Initial Test Score (Internal) is the score in the initial Spanish test based on Duolingo lessons and Initial Test Score (External) is the score in the initial WebCAPE test

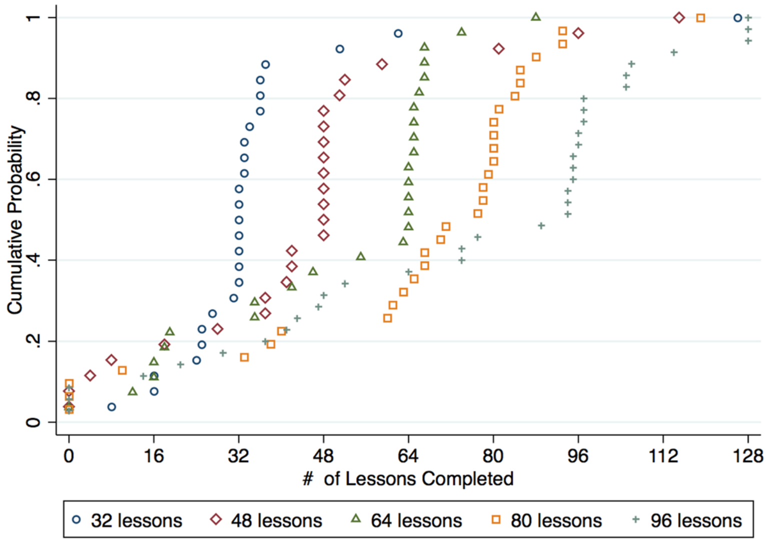

Fig. 2 Lesson Assignments and Number of Lessons Completed. The cumulative distribution of the number of lessons completed in each assignment group for all the subjects in the baseline sample. Number of lessons completed is top coded at 128 for the purposes of this graph. There are two subjects who completed more than 128 lessons. One subject completed 135 lessons and the other subject completed 153 lessons, both in the *96-lessons* group. All the statistics reported in the text and tables and all the tests conducted include these subjects with their actual number of completed lessons

5.1 Effect of lesson assignments on effort

In this subsection, I analyze how random assignment to different number of lessons affects the effort levels of the subjects. Figure 2 depicts the cumulative distribution function of the number of Duolingo lessons completed during the study period separately for all five assignment groups for all the subjects, including those who did not take the final tests. As the figure shows, the median number of lessons completed in each assignment group is in line with the number of assigned lessons. The median participant in the *32-lessons* group completes 32 lessons. The median participants in the *48-lessons*, *64-lessons*, *80-lessons*, and *96-lessons* groups complete 48, 64, 77, and 94 lessons, respectively.[24] The difference between the distributions of the number of completed lessons for any two assignment group is statistically significant (p-value < 0.1 in Mann-Whitney tests and p-value < 0.01 in Kolmogorov-Smirnov tests).

Table 2 shows how lesson assignments affect effort on average for the subjects who completed both of the final tests. The odd-numbered columns do not include any controls whereas even-numbered columns control for initial test scores,

[24] Restricting attention to the subjects who completed both of the final tests reveal similar results. The median participants in the 32-lessons, 48-lessons, 64-lessons, 80-lessons, and 96-lessons groups completed 32, 48, 64, 78, and 95 lessons, respectively.

demographic characteristics, and personality traits. Columns (1)–(4) present the results for the number of Duolingo lessons completed. Column (1) shows that for each additional Duolingo lesson assigned, subjects complete 0.84 more lessons on average (statistically significant at the 1% level). Column (3) depicts the average number of lessons completed in each assignment group. Subjects in the *32-lessons* group complete 30.91 lessons on average, subjects in the *48-lessons* group complete 49.32 lessons on average, and so forth. Columns (5)–(8) present the results for minutes spent on Duolingo, the alternative effort measure. Column (5) shows that for each additional Duolingo lesson assigned, subjects spend 3.22 minutes on Duolingo on average (statistically significant at the 1% level). Column (7) depicts the average time each assignment group spent on Duolingo. Subjects in the *32-lessons* group spend 136.1 minutes on average during the study period. For subjects in the *64-lessons* group, time spent on Duolingo is almost doubled (250.8 minutes, on average). The estimates remain similar after controlling for initial test scores, demographics, and personality traits.[25]

5.2 Effect of effort on test scores

In this subsection, I analyze how study effort within Duolingo affects test performance. To do so, I use the instrumental variables approach where the number of assigned lessons is an instrument for effort. In particular, I estimate the following equation:

$$TestScore_i = \alpha_0 + \alpha_1 \widehat{Effort_i} + X_i \Lambda + \epsilon_i \tag{1}$$

where $TestScore_i$ is either the final internal test score or the final external test score, $Effort_i$ is either the number of lessons completed or minutes spent within Duolingo, and X_i is the set of controls. In the tables, I report both asymptotic p-values and bootstrapped p-values for $Effort_i$. The discussion of significance is based on asymptotic p-values in this section whereas Sect. 5.5.2 discusses significance based on bootstrapped p-values. $\widehat{Effort_i}$ is estimated with the following equation:

$$Effort_i = \beta_0 + \beta_1 NumberofLessonsAssigned_i + X_i \Omega + \nu_i \tag{2}$$

where $NumberofLessonsAssigned_i$ takes values of 32, 48, 64, 80 or 96 depending on subjects' random assignment group. Number of lessons assigned is a valid instrument for effort if it affects the amount of effort exerted (relevance) and affects test scores only through its effect on Duolingo effort (exclusion restriction). As evident from Fig. 2 and Table 2, random assignment of the subjects into lesson groups is successful at affecting the number of lessons completed within Duolingo and minutes spent on Duolingo. Hence, the instrument satisfies the relevance assumption. I discuss the validity of the exclusion restriction in Sect. 5.4.

[25] Repeating the analysis of this table using the number of *new* Duolingo lessons completed and *self-reported* minutes spent results in similar estimates. See Appendix Table B.2 in ESM.

Table 2 Lesson assignment and effort relationship

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Number of lessons completed				Minutes spent on duolingo			
# of Lessons Assigned	0.843***	0.818***			3.222***	3.198***		
	(0.0902)	(0.105)			(0.482)	(0.593)		
48-lessons group			18.41***	20.35**			70.08**	87.45**
			(5.678)	(8.007)			(27.89)	(38.85)
64-lessons group			26.75***	34.93***			114.7***	141.5***
			(4.462)	(6.908)			(28.63)	(36.74)
80-lessons group			39.24***	38.05***			163.2***	169.4***
			(5.310)	(6.181)			(32.27)	(34.47)
96-lessons group			56.80***	57.26***			211.7***	219.1***
			(6.153)	(7.810)			(33.19)	(44.82)
Constant	5.168	− 7.189	30.91***	11.37	41.57	− 34.34	136.1***	60.77
	(5.319)	(51.70)	(2.074)	(53.30)	(28.08)	(282.4)	(13.11)	(285.3)
Observations	120	120	120	120	120	120	120	120
Controls	No	Yes	No	Yes	No	Yes	No	Yes

OLS regressions. Number of Lessons Completed refers to the Duolingo lessons completed over four weeks. Minutes spent on Duolingo refers to the time spent on Duolingo in minutes. # of Lessons Assigned can take values of 32, 48, 64, 80 or 96 depending on subjects' random assignment group. For Columns (3) and (7), Constant represents the omitted group (*32-lessons*). Controls are initial test score in the Duolingo-based test, initial test score in the external test, age, sex, ethnicity dummies, college major, cohort, SAT/ACT math and reading scores, whether SAT/ACT scores are missing, GPA, locus of control score, self control score, growth mindset score, big five personality traits, mother's years of education, father's years of education, annual income, weekly expenditure, target Spanish level in a month, and target Spanish level in a year. Robust standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3 Panel A shows how Duolingo effort affects the performance in Duolingo-based test.[26] Columns (1)–(3) present the findings for the preferred effort measure, number of lessons completed within Duolingo. Column (1) shows that each completed Duolingo lesson increases the internal test score by 0.010 sd (statistically significant at the 5% level). The coefficient remains similar when I control for initial test scores in the internal and external tests in Column (2). As Column (3) shows, each completed Duolingo lesson increases the internal test score by 0.011 sd (statistically significant at the 1% level) after controlling for initial test scores, demographics, and personality traits. Hence, completing 9 Duolingo lessons, which corresponds to approximately 60 minutes of studying, leads to a 0.095 sd increase in the internal test score.[27] Columns (4)–(6) present the results using the minutes spent

[26] Appendix Figure B.2b in ESM left panel displays the distribution of final scores and Appendix Figure B.2c in ESM left panel displays the distribution of improvement in scores across treatment groups for the Duolingo-based test.

[27] Based on a pilot for which I have administrative data on the actual time spent on Duolingo lessons (instead of timestamps), a Duolingo lesson takes 6.7 minutes, on average. Hence, 9 lessons would take approximately 60 minutes.

Table 3 Effort-performance relationship

Panel A	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable	Final score for duolingo-based test (standardized)					
Effort variable:	Number of lessons completed			Minutes spent on duolingo		
Effort	0.0101**	0.00816**	0.0106***	0.00251**	0.00199**	0.00255***
	(0.00492)	(0.00352)	(0.00371)	(0.00128)	(0.000892)	(0.000959)
Asymptotic P-Value	*0.040*	*0.020*	*0.004*	*0.050*	*0.026*	*0.008*
Bootstrap P-Value	*0.042*	*0.010*	*0.018*	*0.046*	*0.010*	*0.020*
Initial Score Internal Test (Std.)		0.739***	0.646***		0.717***	0.610***
		(0.0970)	(0.130)		(0.0939)	(0.136)
Initial Score External Test (Std.)		0.0123	0.0596		0.0665	0.112
		(0.103)	(0.124)		(0.103)	(0.133)
First Stage F-stat	57.75	56.88	38.42	42.85	43.96	26.27
Observations	129	129	129	129	129	129
Panel B						
Dependent variable	Final score for external test (standardized)					
Effort variable	Number of lessons completed			Minutes spent on duolingo		
Effort	0.00756*	0.00428	0.00638**	0.00198*	0.00110	0.00163*
	(0.00447)	(0.00349)	(0.00325)	(0.00120)	(0.000905)	(0.000842)
Asymptotic P-Value	*0.091*	*0.220*	*0.050*	*0.098*	*0.225*	*0.053*
Bootstrap P-Value	*0.096*	*0.226*	*0.102*	*0.092*	*0.226*	*0.104*
Initial Score Internal Test (Std.)		0.421***	0.465***		0.399***	0.437***
		(0.116)	(0.123)		(0.120)	(0.123)
Initial Score External Test (Std.)		0.292**	0.217*		0.328**	0.246*
		(0.123)	(0.129)		(0.127)	(0.134)
First Stage F-stat	87.40	89.78	60.31	44.60	46.72	29.09
Observations	120	120	120	120	120	120
Controls	No	No	Yes	No	No	Yes

2SLS regressions. The dependent variable is the final test score in the Duolingo-based test in standard deviation units for Panel A and the final test score in the external test (WebCAPE) in standard deviation units for Panel B. The instrument is the number of lessons assigned which can take values of 32, 48, 64, 80 or 96 depending on subjects' random assignment group. Number of Lessons Completed refers to the Duolingo lessons completed over four weeks. Minutes spent on Duolingo refers to the time spent on Duolingo in minutes. Initial Score Internal Test (Std.) is the initial score in the Duolingo-based test in standard deviation units and Initial Score External Test (Std.) is the initial score in WebCAPE test. Controls are age, sex, ethnicity dummies, college major, cohort, SAT/ACT math and reading scores, whether SAT/ACT scores are missing, GPA, locus of control score, self control score, growth mindset score, big five personality traits, mother's years of education, father's years of education, annual income, weekly expenditure, target Spanish level in a month and target Spanish level in a year. Robust standard errors are in parentheses. *s are based on asymptotic standard errors. $* \; p < 0.10$, $** \; p < 0.05$, $*** \; p < 0.01$. The asymptotic p-values in italics are based on asymptotic standard errors. The Bootstrap p-values in italics are based on the Wild restricted efficient bootstrap (Davidson and MacKinnon 2010)

on Duolingo during the study as an alternative effort measure. After controlling for initial test scores, demographics, and personality traits (Column (6)), each 60-minute of Duolingo increases the internal test score by 0.153 sd (statistically significant at the 1% level).[28,29]

Table 3 Panel B presents how study effort within Duolingo affects performance in the external test (WebCAPE).[30] Columns (1)–(3) present the findings using the number of lessons completed within Duolingo as the effort measure. Column (1) shows that each completed Duolingo lesson increases the external test score by 0.008 sd (statistically significant at the 10% level). When I control for initial test scores in the internal and external tests in Column (2), the coefficient becomes 0.004 sd (not statistically significant). As Column (3) shows, each completed Duolingo lesson increases the external test score by 0.006 sd (statistically significant at the 5% level) after controlling for initial test scores, demographics, and personality traits. Hence, completing 9 Duolingo lessons, which corresponds to approximately 60 minutes of studying, leads to a 0.057 sd increase in the external test scores. Columns (4)–(6) present the results using the minutes spent on Duolingo during the study as an alternative effort measure. After controlling for initial test scores, demographics, and personality traits (Column (6)), each 60-minute of Duolingo increases the external test score by 0.098 sd (statistically significant at the 10% level).[31,32]

Comparing the effect of effort on *external* test scores (Table 3 Panel B) to the effect of effort on *internal* test scores (Table 3 Panel A) reveals that the effects on the external test are approximately 60% of the effects on the internal test in terms of standard deviations. Given that the internal test is based on Duolingo lessons, stronger and more statistically significant effects for the internal test is expected. However, the existence of the effects for the external test shows that participants in the study indeed learn Spanish skills measured by an external placement test. A simple correlation exercise strengthens this point. I find that final test scores in the internal and external tests are positively correlated with a coefficient of 0.694 (the p-value is less than 0.001). This finding suggests that what is taught in Duolingo is not orthogonal to what is tested in the external test.

How do the estimated effects in this paper compare to the effects reported in the literature for in-person courses? According to estimates in Table 3 Column (3), a one standard deviation increase in number of lessons completed (30.26 lessons)

[28] Appendix Table B.3 in ESM Columns (1)–(2) report the intention-to-treat estimates for Duolingo-based test. Each assigned Duolingo lesson increases the internal test score by 0.008 sd (statistically significant at the 5% level).

[29] The instrumental variables estimates reported in Table 3 Panel A are larger than the corresponding OLS estimates (see Appendix Table B.4 Panel A in ESM). This result suggests that unobserved ability and Duolingo effort are negatively correlated.

[30] Appendix Figure B.2b in ESM right panel displays the distribution of final scores and Appendix Figure B.2c in ESM right panel displays the distribution of improvement in scores across treatment groups for the external test.

[31] Appendix Table B.3 in ESM Columns (3)–(4) report the intention-to-treat estimates for WebCAPE. Each assigned Duolingo lesson increases the external test score by 0.005 sd (statistically significant at the 10% level).

[32] The instrumental variables estimates reported in Table 3 Panel B are larger than the corresponding OLS estimates (see Appendix Table B.4 Panel B in ESM).

increases the internal test scores by 0.32 sd (statistically significant at the 1% level) and the external test scores by 0.19 sd (statistically significant at the 5% level). According to estimates in Table 3 Column (6), a one standard deviation increase in time spent (144.24 minutes) increases the internal test scores by 0.37 sd (statistically significant at the 1% level) and the external test scores by 0.24 sd (statistically significant at the 10% level). These effects are comparable to the effects reported in the literature for in-person (non-language) courses (0.20 sd in Eren and Henderson (2011), 0.25 sd in Bonesrønning and Opstad (2012), and 0.14 sd in Bonesrønning and Opstad (2015)).

How much effort does a student need to exert in Duolingo to advance to another semester? Since WebCAPE is used as a placement test in many colleges, I can use the suggested cutoffs for different semesters to answer this question. WebCAPE guidelines state that a student with a score of 270 points could be placed in Semester 2 of a 4-semester Spanish course.[33] According to the estimates in Table 3 Panel B Column (3), each additional lesson increases test scores by 0.785 points (statistically significant at 5% level).[34] Similarly, according to estimates in Table 3 Panel B Column (6), each additional minute increases test scores by 0.201 points (statistically significant at 10% level). If we take these estimates at face value and linearly extrapolate, back-of-the-envelope calculations suggest that a student who does not know any Spanish needs to complete 344 Duolingo lessons (270/0.785 based on Table 3 Panel B Column (3)) or spend 22 hours on Duolingo (270/0.201 based on Table 3 Panel B Column (6)) to be placed in Semester 2. The former corresponds roughly to 38 hours.[35] These amounts are comparable to or less than what students spend in a classroom in a semester (roughly 45 hours of class time without including time spent studying outside of class for a semester) and suggest that effort in Duolingo can be as effective as effort in in-person classrooms.[36] However, this conclusion is suggestive due to two factors. First, the maximum number of assigned lessons is 96 and the maximum number of completed lessons is 153. This exercise assumes linear extrapolation for 344 lessons, which is far out of the sample. If the returns to effort are diminishing, this exercise would overestimate the effectiveness of effort in Duolingo. Second, since the confidence intervals for the estimated effect for the external test are quite large, it is possible that the actual effectiveness of Duolingo effort is smaller than the one reported here.

[33] A student with a score of 346 points could be placed in Semester 3 and a student with a score of 428 points could be placed in Semester 4 of a 4-semester Spanish course.

[34] Notice that this discussion in the text is based on points and the estimates in the table are based on standard deviations.

[35] Duolingo reports that an average lesson takes 5 to 10 minutes. Based on a pilot for which I have administrative data on the actual time spent on Duolingo lessons (instead of timestamps), a lesson takes 6.7 minutes, on average. Using this estimate, I find that 344 Duolingo lessons will take approximately 38 hours ($\frac{344*6.7}{60}$).

[36] A semester corresponds to 15 weeks and 3 hours of instruction per week per class for most of the schools in my sample.

5.3 Heterogenous effects of effort on test scores

In this subsection, I explore whether the returns to effort are heterogenous by college GPA. Effort might not translate into performance at the same rate for everyone. It is important to understand whether returns to effort in online platforms differ for higher-achieving and lower-achieving students to make better policy recommendations. To do so, I estimate the following equation:

$$\begin{aligned} FinalTestScore_i = \alpha_0 + \alpha_1 \widehat{Effort_i} + \alpha_2 HighGPA_i \\ + \alpha_3 \widehat{EffortHighGPA_i} + X_i \Lambda + \epsilon_i \end{aligned} \tag{3}$$

where $HighGPA_i$ is a dummy variable which is equal to 1 if an individual's GPA is above the median GPA and 0 otherwise. $\widehat{Effort_i}$ and $\widehat{EffortHighGPA_i}$ are estimated by the following equations:

$$\begin{aligned} Effort_i &= \beta_0 + \beta_1 NumberofLessonsAssigned_i + \beta_2 HighGPA_i \\ &\quad + \beta_3 NumberofLessonsAssigned_i * HighGPA_i + X_i \Omega + \nu_i \\ EffortHighGPA_i &= \gamma_0 + \gamma_1 NumberofLessonsAssigned_i + \gamma_2 HighGPA_i \\ &\quad + \gamma_3 NumberofLessonsAssigned_i * HighGPA_i + X_i \Gamma + \zeta_i \end{aligned} \tag{4}$$

Table 4 Columns (1) and (2) present the results for the internal test and two effort measures: number of lessons completed and minutes spent on Duolingo, respectively. I find that Duolingo is more effective for students who have a higher college GPA. In particular, the effects are large and statistically significant at the 1% level for students with above-median college GPA but small and not statistically significant for students who have a below-median college GPA. Table 4 Columns (3) and (4) explore heterogeneity in returns to effort for the external test. I again observe heterogeneity in returns to effort based on GPA. The effects are large and statistically significant at the 5%-10% level for above-median students, but small and not statistically significant for below-median students.[37] Similar to these findings, Bettinger et al. (2017) find that effectiveness of online courses positively correlates with prior GPA.

5.4 Exclusion restriction

Section 5.2 presents the causal effect of Duolingo effort on test scores under the assumption that the instrument satisfies the exclusion restriction. In this section, I explore the validity of this assumption.

First, the exclusion restriction of the instrumental variables estimation will fail if participants use other online and offline platforms to learn Spanish at different

[37] Using a continuous GPA measure also results in similar findings. A one point increase in GPA interacted with the number of lessons completed with Duolingo increases the internal (external) test scores by 0.0367 sd (0.0274 sd) and a one point increase in GPA interacted with minutes spent on Duolingo increases the internal (external) test scores by 0.008 sd (0.007 sd).

Table 4 Effort-performance relationship: heterogeneity by college GPA

Dependent variable	(1)	(2)	(3)	(4)
	Final score for internal test (Std)		Final score for external test (std)	
Effort variable	# of lessons completed	Time spent	# of lessons completed	Time spent
Effort	0.0008	0.0004	0.0002	0.0002
	(0.0048)	(0.0011)	(0.0049)	(0.0011)
Asymptotic P-Value	*0.875*	*0.748*	*0.972*	*0.872*
Bootstrap P-Value	*0.861*	*0.753*	*0.939*	*0.897*
Effort*High GPA	0.0199***	0.0061***	0.0138*	0.0042**
	(0.0073)	(0.0021)	(0.0077)	(0.0021)
Asymptotic P-Value	*0.007*	*0.004*	*0.073*	*0.042*
Bootstrap P-Value	*0.022*	*0.014*	*0.114*	*0.052*
Initial Score Internal Test (Std.)	0.6267***	0.5811***	0.4555***	0.4260***
	(0.1149)	(0.1298)	(0.1239)	(0.1277)
Initial Score External Test (Std.)	0.1321	0.2261	0.2568**	0.3192**
	(0.1186)	(0.1398)	(0.1286)	(0.1408)
Observations	129	129	120	120
Controls	Yes	Yes	Yes	Yes

2SLS regressions. The dependent variable is the final test score in the Duolingo-based test in standard deviation units for Columns (1)–(2) and the final test score in the external test (WebCAPE) in standard deviation units for Columns (3)–(4). Effort in Columns (1) and (3) is Number of Lessons Completed and Effort in Columns (2) and (4) is Minutes spent on Duolingo. High GPA is a dummy variable which is equal to 1 if an individual's GPA is above the median GPA and 0 otherwise. The instruments are the number of lessons assigned and number of lessons assigned interacted with High GPA. Controls are initial test score in the Duolingo-based test, initial test score in the external test (WebCAPE), age, sex, ethnicity dummies, college major, cohort, SAT/ACT math and reading scores, whether SAT/ACT scores are missing, GPA, locus of control score, self control score, growth mindset score, big five personality traits, mother's years of education, father's years of education, annual income, weekly expenditure, target Spanish level in a month, and target Spanish level in a year. Robust standard errors are in parentheses. *s are based on asymptotic standard errors. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The asymptotic p-values in italics are based on asymptotic standard errors. The Bootstrap p-values in italics are based on the Wild restricted efficient bootstrap (Davidson and MacKinnon 2010)

amounts based on their lesson assignments. The participants in the *32-lessons* group might use outside resources more if they think that Duolingo lessons are not enough or might use outside resources less if they become demotivated compared to the participants in the *96-lessons* group. Such behavior would violate the exclusion restriction. To check this violation, the final survey asks participants whether they have used any online or offline tools to learn Spanish during the study period and how much time they spent studying Spanish with means other than Duolingo in the last 7 days of the study. Table 5 Column (1) (Column (2)) looks at whether the number of lessons assigned (assignment groups) affects subjects' usage of other tools. Although approximately 5% of the subjects report using other tools, the instrument does not affect the usage of other tools. Column (3) (Column (4)) presents the effect of the number of lessons assigned

Table 5 Lesson Assignment and Other Outcomes

	(1)	(2)	(3)	(4)	(5)	(6)
	Other tools		Minutes spent outside of duolingo		Cheating in the final exam	
# of Lessons Assigned	− 0.00058		0.621		0.0014	
	(0.00094)		(0.615)		(0.0009)	
48-lessons group		− 0.0318		− 45.39		0.0198
		(0.107)		(51.95)		(0.0376)
64-lessons group		− 0.139		− 44.06		0.0117
		(0.0938)		(42.78)		(0.0387)
80-lessons group		− 0.0373		− 13.01		0.0686
		(0.0825)		(26.83)		(0.0483)
96-lessons group		− 0.0544		20.48		0.0853
		(0.0760)		(28.51)		(0.0563)
Constant	0.354	0.351	429.2	442.1	0.492	0.544
	(0.619)	(0.598)	(292.4)	(306.8)	(0.399)	(0.404)
Mean of Dependent Variable	0.0508		17.38		0.0254	
Median of Dependent Variable	0		0		0	
Observations	118		117		118	
Controls	Yes		Yes		Yes	

OLS regressions. Other Tools is a dummy variable which takes a value of one if a subject reports using any online or offline tools to learn Spanish other than Duolingo during the study period. Minutes Spent Outside of Duolingo refers to (self-reported) time spent on studying Spanish with means other than Duolingo in the last seven days. Cheating in the Final Exam is a dummy variable which takes a value of one if a subject reports using sources other than his/her own knowledge while taking the final Spanish tests. Number of Lessons Assigned can take values of 32, 48, 64, 80 or 96 depending on subjects' random assignment group. For Columns (2), (4), and (6), Constant represents the omitted group (*32-lessons*). For Columns (2), (4), and (6), I cannot reject the null hypothesis of all coefficients being equal to zero. Controls are initial test score in the Duolingo-based test, initial test score in the external test (WebCAPE), age, sex, ethnicity dummies, college major, cohort, SAT/ACT math and reading scores, whether SAT/ACT scores are missing, GPA, locus of control score, self control score, growth mindset score, big five personality traits, mother's years of education, father's years of education, annual income, weekly expenditure, target Spanish level in a month and target Spanish level in a year. Robust standard errors are in parentheses. $* p < 0.10$, $** p < 0.05$, $*** p < 0.01$

(assignment groups) on how many minutes subjects spent on studying Spanish with means other than Duolingo. Subjects report spending around 17 minutes on average outside of Duolingo to learn Spanish but the instrument does not affect this variable. Appendix Table B.5 in ESM repeats the main analysis in Sect. 5.2 for the subsample of participants who report using no other tools. The estimates are similar to the ones shown in Table 3. However, we should be careful about the interpretation of these results since the data on both the usage of other platforms and time spent on other platforms are self-reported and are likely to suffer from reporting error and experimenter demand effects.

Second, the instrument might impact subjects' test scores by changing their test-taking strategies due to experimental demand effects. To check if such a violation of the exclusion restriction occurred, the final survey asks subjects whether they used sources other than their own knowledge while taking the final Spanish tests. 2.5% of them reported using outside sources but there is no differential impact based on the instrument as Table 5 Columns (5) and (6) show. Hence, the instrument does not statistically significantly change exam-taking behavior.[38] Appendix Table B.6 in ESM repeats the main analysis in Sect. 5.2 for the subsample of the participants who report using no other sources while taking the final Spanish tests. The estimates are similar to the ones shown in Table 3.

5.5 Robustness

5.5.1 Attrition

In this subsection, I present a placebo check to address differential attrition based on initial test scores, perform a bounding exercise for the treatment effects, and compute inverse probability-weighted estimates of treatment effects.

To check whether there is differential attrition related to the initial test scores, I estimate Equation 1 by using the initial test scores as the dependent variable and restricting the sample to the end sample. Appendix Table B.7 in ESM shows the results of this placebo check. The estimated coefficients are quite small compared to the main effects and far from significant for both internal and external test scores. This result shows that there is no difference in initial test scores in the end sample.

To estimate the bounds for the treatment effects, I use the approach described in Horowitz and Manski (2000). This approach does not make any distributional assumptions and involves imputing the missing values of the dependent variable with the maximum or the minimum values in the support of the dependent variable. Appendix Table B.8 in ESM presents the results of this bounding exercise. Panel A shows the results for the internal test scores. I find that each completed Duolingo lesson increases internal test scores by 0.009 sd (statistically significant at the 10% level) when I replace the missing final test scores with the minimum final test score and by 0.016 (statistically significant at the 5% level) when I replace the missing final test scores with the maximum final test score. Hence, the exercise turns out be informative for internal test scores. Panel B shows the results for the external test scores. The lower bound is 0.004 sd (not statistically significant) and the upper bound is 0.014 sd (statistically significant at the 5% level). The bounding exercise in this case turns out not to be informative since the minimum bound includes zero in its 90% confidence interval.[39]

[38] Additionally, we might worry that the instrument affects the usage frequency of the platform and usage frequency has a direct effect on test performance. To explore this concern, I check how the number of days active in the platform differs across lesson groups. Although there are some statistical differences, the differences are not monotonic and somewhat different in each week.

[39] Doing a trimming exercise in the spirit of Lee (2009) bounds results in similar findings with slightly tighter bounds. In particular, the exercise is informative for the internal test scores (the lower bound is 0.008 sd (statistically significant at the 5% level) and the upper bound is 0.011 sd (statistically significant at the 1% level)), but not for the external test scores (the lower bound is 0.005 sd (not statistically significant) and upper bound is 0.008 sd (statistically significant at the 1% level)).

To estimate inverse-probability weighted treatment effects, I predict the probability of taking the final test based on the initial test scores and use the inverse of the predicted probability as weights in the instrumental variables regression using a one-step GMM estimation procedure. Appendix Table B.9 in ESM presents the results. The estimates are very similar to that of Table 3.

5.5.2 Small sample size

Statistically significant results can be misleading in small sample settings. To shed some light on how important this issue is for my estimates, I follow the procedures in Gelman and Carlin (2014) who propose a set of statistical calculations about what could happen under hypothetical replications of the study. They combine information from the current experiment (sample size, standard error of the estimate, and degrees of freedom) with external information from the literature (on true effect size) to calculate the power, sign error rate, and exaggeration ratio for the effect that would be observed in a hypothetical replication study with a design like the one used in the original study. The power is defined as the probability that the replicated estimate is larger than the critical value. The sign error rate refers to the probability that the replicated estimate has the incorrect sign if it is statistically significantly different from zero. The exaggeration ratio is defined as the expected overestimation of the magnitude of the effect if it is statistically significantly different from zero. Following Gelman and Carlin (2014), I have calculated the power, sign error rate and exaggeration ratio for a range of possible true effect sizes from the literature. I calculate these statistics for effect sizes of 0.1 sd, 0.2 sd and 0.3 sd since the reported effect sizes in the literature range from 0.14 to 0.25 sd for returns to effort in in-person classes. Appendix Table B.10 in ESM displays the results. For example, if the true effect for the internal test scores is 0.2 sd, power is 0.422, and there is less than 0.001% chance that the estimated effect has the wrong sign and the estimated effect is 1.53 times too high (in expectation) conditional on the estimate being statistically significant. The results for external test scores are quite similar.

Additionally, asymptotic inference might be unreliable in the case of small samples. In the case of small samples, bootstrapping can provide more accurate inferences since it does not require distributional assumptions. Hence, I report the Wild restricted efficient bootstrap p-values (Davidson and MacKinnon 2010) in addition to asymptotic p-values in Tables 3 and 4. For internal tests, all the results continue to be statistically significant with a bootstrap p-value less than 0.05 in all specifications for both effort measures. However, the bootstrap procedure suggests that the estimates for the external test scores are noisier than what the asymptotic inference suggests. I also conduct randomization inference (Athey and Imbens 2017) for the intent-to-treat estimates of Appendix Table B.3. in ESM In particular, I create placebo treatment groups by randomly assigning the subjects into treatment groups while preserving the original number of subjects in each group. I then calculate the effect of the placebo number of lessons assigned on test scores. I repeat this procedure 500 times. I then look at the probability that the absolute value of the placebo effect sizes are larger than the original effect size (p-value). P-values based on randomization inference are reported in brackets in Appendix Table B.3 in ESM and are similar to the ones based on asymptotic inference.

5.5.3 Alternative effort measures and alternative instruments

In my main specification, I use all the lessons participants completed when calculating the "number of lessons completed" variable. This measure includes new as well as repeated lessons (lessons completed more than once). In the experiment, 69% of the participants did not complete any repeated lessons and 20% of them completed between 1 and 5 repeated lessons. Since returns to new and repeated lessons might be different, I repeat the main analysis using only new lessons completed in Appendix Table B.11 in ESM Columns (1)–(3). The results are very similar to that of Table 3.

In my main specification, I use administrative data on timestamps for all lessons to calculate the "time spent" variable. Since timestamps only indicate when a lesson was completed but do not provide information about when a lesson was started, I create a proxy for the "time spent" variable as discussed in Sect. 4.1. As an alternative to this proxy "time spent" measure, I repeat the main analysis using self-reported minutes of study on Duolingo based on the final survey. Appendix Table B.11 in ESM Columns (4)–(6) report the results. The results are very similar to that of Table 3 but the first-stage is weaker for this self-reported measure of effort.

In this experiment, I randomly assign participants to different treatment groups to create exogenous variation in effort. The only variation across treatment groups is the number of lessons assigned. Hence, I convert treatment group dummies to a single instrument, the number of lessons assigned, in my main specification. An alternative to this approach is to use treatment group dummies as instruments. Appendix Table B.12 in ESM shows the results of this alternative approach. The estimates are similar to that of Table 3 but the first-stage F statistics are substantially smaller.

6 Conclusion

This paper examines returns to effort in an online language learning platform. Although distance learning has increased in prevalence, not much is known about how effort affects performance in these settings. High dropout rates in these settings and endogeneity of study effort cause estimation challenges. To tackle these challenges, I design an experiment in which I randomly assign students into groups with varying degrees of lesson assignments. The assignments manipulate students' effort choices. I then estimate how effort translates into performance. I find that effort increases performance in both the internal and external tests. The effect sizes are comparable to the returns to effort estimates from in-person classrooms. Furthermore, students with higher GPAs have higher returns to effort in the online platform.

We should be careful about the limitations of this study and generalizability of its results. One of the limitations of this study is its lack of control on the use of other online and offline platforms to learn Spanish. Another limitation is the small sample size. Furthermore, the positive effects of effort on performance estimated in this study is for a beginner-level course and for a population of intrinsically motivated individuals. The effects might be different for an upper-level course or for a required course; more research on the topic is needed.

All in all, this paper contributes to the literature by documenting substantial positive effects of effort on achievement in a technology-assisted remote learning setting. The heterogeneous returns by GPA suggest that these settings can be especially beneficial for higher achieving students. The findings of this study indicate remote settings may be a viable alternative to in-person environments conditional on students' exerting effort in a timely manner in these settings. Given low levels of effort in distance learning environments, policymakers and course designers might want to improve students' time management skills to reap the benefits of remote settings. However, Oreopoulos et al. (2019) show that this task is difficult.

Acknowledgements I thank Sandro Ambuehl, Douglas Bernheim, Eric Bettinger, Raj Chetty, Pascaline Dupas, Sarah Eichmeyer, Matthew Gentzkow, Caroline Hoxby, Susanna Loeb, Daniel Martin, Muriel Niederle, Odyssia Ng, Imran Rasul, Georg Weizsäcker, David Yang, and participants at Western Economic Association International (WEAI) 2019 and Economic Science Association (ESA) 2019 for helpful comments and discussions. This research was conducted with the support of Russell Sage Foundation (Grant No. 98-16-14). This research is approved by Stanford Panel on Non-Medical Human Subjects (IRB protocol # 36512). This experiment is registered at the AEA RCT registry (ID 0001775) as Stage 1 of a two-stage experiment. The results from Stage 2 are available in a companion paper.

References

Angrist, J., & Lavy, V. (2009). The effects of high stakes high school achievement awards: Evidence from a randomized trial. *American Economic Review*, *99*(4), 301–331.

Athey, S., & Imbens, G. (2017). The econometrics of randomized experiments. *Handbook of Economic Field Experiments*, *1*, 73–140.

Banerjee, A., Cole, S., Duflo, E., & Linden, L. (2007). Remedying education: Evidence from two randomized experiments in India. *The Quarterly Journal of Economics,* (pp. 1235–1264).

Barrow, L., Markman, L., & Rouse, C. E. (2009). Technology's edge: The educational benefits of computer-aided instruction. *American Economic Journal: Economic Policy*, *1*(1), 52–74.

Bawa, P. (2016). Retention in online courses: Exploring issues and solutions—a literature review. *SAGE Open*.

Bettinger, E. P., Fox, L., Loeb, S., & Taylor, E. S. (2017). Virtual classrooms: How online college courses affect student success. *American Economic Review*, *107*(9), 2855–2875.

Betts, J.R. (1996). The role of homework in improving school quality. Discussion Paper 96–16. Department of Economics, UCSD.

Bonesrønning, H., & Opstad, L. (2012). How much is students' college performance affected by quantity of study? *International Review of Economics Education*, *11*, 46–63.

Bonesrønning, H., & Opstad, L. (2015). Can student effort be manipulated? Does it matter? *Applied Economics*.

Chen, J., & Lin, T.-F. (2008). Class attendance and exam performance: A randomized experiment. *The Journal of Economic Education*, *39*(3), 213–227.

Chevalier, A., Dolton, P., & Lührmann, M. (2018). Making it count: Incentives. Student effort and Performance. *Journal of Royal Statistical Society*, *181*, 323–349.

Clark, D., Gill, D., Prowse, V. & Rush, M. (2017). Using goals to motivate college students: Theory and evidence from field experiments. *NBER Working Paper 23638*.

Davidson, R., & MacKinnon, J. G. (2010). Wild bootstrap tests for IV regression. *Journal of Business & Economic Statistics*, *28*(1), 128–144.

De Fraja, G., Oliveira, T., & Zanchi, L. (2010). Must try harder: Evaluating the role of effort in educational attainment. *The Review of Economics and Statistics*, *92*(3), 577–597.

Deming, D. J., Goldin, C., Katz, L. F., & Yuchtman, N. (2015). Can online learning bend the higher education cost curve? *American Economic Review: Paper & Proceedings*, *105*(5), 496–501.

Dobkin, C., Gil, R., & Marion, J. (2010). Skipping class in college and exam performance: Evidence from a regression discontinuity classroom experiment. *Economics of Education Review*, *29*, 566–575.

Eren, O., & Henderson, D. J. (2008). The impact of homework on student achievement. *Econometrics Journal*, *11*, 326–348.

Eren, O., & Henderson, D. J. (2011). Are we wasting our children's time by giving them more homework? *Economics of Education Review*, *30*, 950–961.

Escueta, M., Nickow, A. J., Oreopoulos, P., & Quan, V. (forthcoming). Upgrading education with technology: Insights from experimental research. *Journal of Economic Literature*.

Fryer, R. G. (2011). Financial incentives and student achievement: Evidence from randomized trials. *The Quarterly Journal of Economics*.

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and Individual Differences*, *7*(3), 385–400.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (Sign) and type M (Magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651.

Ghanem, D., Hirshleifer, S. R. & Ortiz-Becerra, K. (2019). Testing attrition bias in field experiments. *Working Paper*.

Glewwe, P., & Muralidharan, K. (2016). Improving education outcomes in developing countries: Evidence, knowledge gaps, and policy implications. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the economics of education* (Vol. 5). London: Elsevier.

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist*, *54*, 493–503.

Goodman, J., Melkers, J., & Pallais, A. (2019). Can online delivery increase access to education? *Journal of Labor Economics*, *37*(1), 1–34.

Grodner, A., & Rupp, N. G. (2013). The role of homework in student learning outcomes: Evidence from a field experiment. *The Journal of Economic Education*, *44*(2), 93–109.

Hirshleifer, S. R. (2016). Incentives for effort or outputs? A field experiment to improve student performance. *Working Paper*.

Horowitz, J. L., & Manski, C. F. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association*, *95*(449), 77–84.

Jordan, K. (2015). MOOC completion rates: The data. http://www.katyjordan.com/MOOCproject.html.

Kamenica, E. (2012). Behavioral economics and psychology of incentives. *The Annual Review of Economics*.

Kizilcec, R. F. & Brooks, C. (2017). Diverse big data and randomized field experiments in massive open online courses. In A. Wise C. Lang, G. Siemens & D. Gasevic, (eds.), *Handbook on learning analytics & educational data mining* (pp. 211–222).

Kizilcec, R. F., Brooks, C., & Halawa, S. (2015). Attrition and achievement gaps in online learning. In *Proceedings of the second ACM conference on learning@scale* (pp. 57–66).

Krohn, G. A., & O'Connor, C. M. (2005). Student effort and performance over the semester. *The Journal of Economic Education*, *36*(1), 3–28.

Kuehn, Z. & Landeras, P. (2012). Study time and scholarly achievement in PISA. *MPRA Paper No. 49033*.

Lai, F., Luo, R., Zhang, L., Huang, X., & Rozelle, S. (2015). Does computer-assisted learning improve learning outcomes? Evidence from a randomized experiment in migrant schools in Beijing. *Economics of Education Review*, *47*, 34–48.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies*, *76*(3), 1071–1102.

Linden, L. (2008). Complement or substitute? The effect of technology on student achievement in india. *Working Paper*.

Metcalfe, R., Burgess, S., & Proud, S. (2019). Students' effort and educational achievement: Using the timing of the World Cup to vary the value of leisure. *Journal of Public Economics*.

Muralidharan, K., Singh, A., & Ganimian, A. J. (2019). Disrupting education? Experimental evidence on technology-aided instruction in India. *American Economic Review*, *109*(4), 1426–1460.

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology*, *15*(3), 263–280.

Oreopoulos, P., Patterson, R. W., Petronijevic, U., & Pope, N. G. (2019). When studying and nudging don't go as planned: Unsuccessful attempts to help traditional and online college students. *NBER Working Paper 25036*.

Owen, J. D. (1995). *Why our kids don't study? An economist's perspective*. Baltimore: The John Hopkins University Press.

Paunesku, D., Walton, G. M., Romero, C., Smith, E. N., Yeager, D. S., & Dweck, C. S. (2015). Mind-Set Interventions are a scalable treatment for academic underachievement. *Psychological Science*, *26*, 784–793.

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the big five inventory in English and German. *Journal of Research in Personality*, *41*(1), 203–212.

Renz, J., Hoffman, D., Staubitz, T. & Meinel, C. (2016). Using A/B testing in MOOC environments Using A/B testing in MOOC environments Using A/B testing in MOOC environments using A/B testing in MOOC environments. In *Proceedings of the sixth international conference on learning analytics & knowledge*.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied*, *80*, 1–28.

Seaman, J. E., Ellaine Allen, I., & Seaman, J. (2018). Grade increase: Tracking distance education in the United States. Technical Report, Babson Survey Research Group.

Stinebrickner, R., & Stinebrickner, T. R. (2004). Time-use and college outcomes. *Journal of Econometrics*, *121*, 243–269.

Stinebrickner, R., & Stinebrickner, T. R. (2008). The causal effect of studying on academic performance. *The B.E. Journal of Economic Analysis and Policy*.

Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality*, *72*(2), 271–324.

Vesselinov, R., & Grego, J. (2012). Duolingo effectiveness study. *Working Paper*.