ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 21-400

Effects of Perceived Productivity on Study Effort: Evidence from a Field Experiment

Fulya Y Ersoy
Loyola Marymount University

How does the perceived relationship between effort and achievement affect effort? To answer this question, I conduct a field experiment with a popular online learning platform. I exogenously manipulate students' beliefs about returns to effort by assigning them to different information treatments, each of which provides factual information. Students update their beliefs towards the information provided and change their study effort in the same direction with the shifts in their beliefs. This result shows that students' beliefs about the returns to effort is an important component of their human capital accumulation and low-cost information interventions can influence these beliefs.

VERSION: May 2021

Suggested citation: Ersoy, Fulya Y. (2021). Effects of Perceived Productivity on Study Effort: Evidence from a Field Experiment. (EdWorkingPaper: 21-400). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/ewbe-ah97

Effects of Perceived Productivity on Study Effort: Evidence from a Field Experiment

Fulya Ersoy*

April 2, 2021

Abstract

How does the perceived relationship between effort and achievement affect effort? To answer this question, I conduct a field experiment with a popular online learning platform. I exogenously manipulate students' beliefs about returns to effort by assigning them to different information treatments, each of which provides factual information. Students update their beliefs towards the information provided and change their study effort in the same direction with the shifts in their beliefs. This result shows that students' beliefs about the returns to effort is an important component of their human capital accumulation and low-cost information interventions can influence these beliefs.

***JEL* Codes:** I26, D03, C93

Keywords: Effort, Beliefs, Returns to Effort, Field Experiment

*Contact: Department of Economics, Loyola Marymount University, 1 LMU Drive, Los Angeles, CA 90045, fulya.ersoy@lmu.edu. I thank Sandro Ambuehl, Douglas Bernheim, Jaime Arellano-Bover, Eric Bettinger, Raj Chetty, Pascaline Dupas, Carol Dweck, Sarah Eichmeyer, Matthew Gentzkow, Caroline Hoxby, David Laibson, Yucheng Liang, Susanna Loeb, Christos Makridis, Joshua Miller, Odyssia Ng, Muriel Niederle, Oriol Pons Benaiges, Imran Rasul, Alvin Roth, Sally Sadoff, Heather Sarsons, Charlie Sprenger, Emmanuel Vespa, Georg Weizsäcker, David Yang, workshop participants at Harvard, Georgia State and UC San Diego, seminar participants at Bilkent University, Loyola Marymount University, Stanford University, UC Santa Barbara, University of New South Wales, University of Nottingham, BABEEW 2017, ESA 2017, Workshop on Subjective Expectations 2020, ASSA 2021, and AEFP 2021 for helpful comments and discussions. I appreciate the generous support of the Institute for Research in the Social Sciences (IRiSS) at Stanford University, Leonard W. Ely and Shirley R. Ely Graduate Student Fellowship through a grant to the Stanford Institute for Economic Policy Research (SIEPR), Russell Sage Foundation, Shultz Graduate Student Fellowship in Economic Policy through a grant to the SIEPR, and Stanford Graduate Research Opportunities Fund. This research is approved by Stanford Panel on Non-Medical Human Subjects (IRB protocol # 36512). This experiment is registered at AEA RCT registry (ID 0001775).

1 Introduction

Study effort is one of the most important determinants of academic achievement (Costrell, 1994; Stinebrickner and Stinebrickner, 2008; De Fraja, Oliveira and Zanchi, 2010). Hence, it is important to understand how students make their effort decisions. The literature on the determinants of effort documents that immediate financial rewards (Fryer, 2011; Bettinger, 2012; Barrow and Rouse, 2018; Hirshleifer, 2016; Levitt, List, Neckermann and Sadoff, 2016), task-based goals (Clark, Gill, Prowse and Rush, 2020), and cultural and social norms (Bishop, 2006; Figlio, Giuliano, Ozek and Spaienza, 2019; Gneezy, List, Livingston, Sadoff, Qin and Xu, 2019; Burstyn, Egorov and Jensen, 2019) affect students' effort decisions. However, we still have only a limited understanding of these decisions. One critical yet unexplored determinant of study effort is students' perceptions about the effort-achievement relationship. Theoretically, the effect of perceived returns to effort on effort is ambiguous and depends on the interplay between the substitution and income effects.[1]

The role of perceptions about returns to effort on study effort has been difficult to investigate due to four challenges. First, students' beliefs regarding the returns to effort are rarely measured.[2] Second, these beliefs are endogenous since they correlate with unobservable factors that affect effort choices. Third, study effort is hard to observe due to its multidimensional nature. Fourth, measures of achievement are often weakly correlated with effort since these measures combine many types of human capital investments. In this study, I overcome these challenges by conducting a framed field experiment in a setting where effort is observable and measurable.[3] In this experiment, I measure students beliefs about the returns to effort repeatedly, I exogenously and successfully manipulate these beliefs through information treatments, and I design a performance test that closely reflects the material. In this setting, I find that students' perceptions about returns to effort affect their choice of effort. In particular, students' effort move in the same direction with the change in their beliefs.

[1] For example, in the case of a decrease in the perceived effectiveness of effort, the substitution effect would make students consume more leisure since the price of leisure is lower, whereas the income effect would make students consume less leisure since their time endowment is less valuable and leisure is a normal good.

[2] This lack of measurement is not restricted to students' beliefs. Workers' beliefs about returns to effort are also not measured. Even old, large and well-recognized surveys, such as Current Population Survey in the US and Labor Force Survey in Canada, do not measure the beliefs about the effort-rewards relationship.

[3] A framed field experiment is structured in the natural environment of the students and with the natural tasks but students are aware that they are taking part in an experiment (Harrison and List, 2004).

To demonstrate the role of students' beliefs about returns to effort on their effort decisions, I conduct the experiment with an online language learning platform, Duolingo. Duolingo is a popular platform with more than 300 million users worldwide and there are many other mobile learning apps that share features with Duolingo.[4] Mobile e-learning is a growing industry. The mobile e-learning market was sized at approximately 19.5 billion U.S. dollars in 2019 and it is forecasted to reach 48.5 billion U.S. dollars by 2026.[5] The biggest challenges these learning apps face are low persistence and low levels of effort (see Bawa, 2016, for a literature review). Hence, understanding how students' beliefs about returns to effort shape their drop-out and effort decisions in these platforms is important.

Duolingo provides a clean setting to study the effects of perceived returns to effort on effort. Duolingo implements a task-based effort measure, number of lessons completed, which is observable to the researcher whereas effort in most educational settings is multi-dimensional and hard to measure. These lessons are the only tasks available to students on Duolingo and students complete the lessons in the exact same sequence. These features make Duolingo environment highly standardized. Furthermore, absence of parents, teachers, and peers in Duolingo setting allows for clear identification of the effects of beliefs on returns to effort on choices of effort. Hence, Duolingo provides both a naturalistic and a controlled setting to study the role of beliefs on study effort.

To explore how students' perceptions about returns to effort affect their effort decisions, I recruit college students who want to learn Spanish online. First, students answer questions about their demographics, personality traits, and educational background. Then, I measure students' initial Spanish knowledge with a test that is based on the contents of Duolingo. After I elicit students' baseline beliefs about how Duolingo lessons affect test performance in an incentivized manner, I randomly assign the students into one of the four treatments to generate exogenous variation in these beliefs. Treatments are based on a previous experiment with participants from the same subject pool (Ersoy, 2020) in which I manipulate Duolingo effort by assigning participants different numbers of lessons to measure the causal returns to effort. In the *No Information* treatment, students do not receive any information about the effort-performance relationship. In the *Average Information* treatment, students learn about the average causal relationship for the participants

[4]The total number of learning app downloads in 2017 stands at 178 billion worldwide. Retrieved from https://elearningindustry.com/big-elearning-market-role-mobile-apps.

[5]Source: https://www.statista.com/statistics/1130331/e-learning-market-size-segment-worldwide/.

of the previous experiment. In the *Low* (*High*) treatment, students receive anecdotal information based on a single student from the previous experiment whose returns were unusually low (high) and whose pre-test score was similar to the score of the recipient of the information. After students are assigned into the treatments, they study Spanish as much as they want for four weeks while I track their effort (number of lessons completed and time spent) on Duolingo. I also continue to elicit their beliefs about the effort-performance relationship on a weekly basis. At the end of the four weeks, students take a test that measures their final Spanish knowledge and a final survey that asks questions on the usage of other platforms and questions about potential confounders.

In this setting, students' perceptions about returns to effort are inaccurate, and the majority of them overestimate returns to effort at the baseline. Information treatments move students' beliefs about returns to effort towards the information provided, regardless of the type of the information. When individuals receive information about the average causal returns to effort based on the previous experiment, their beliefs move towards the information (with the average individual becoming less optimistic). When individuals receive anecdotal information about the effort-performance relationship of a single previous participant, they also update their beliefs towards the information (with the average individual becoming less optimistic in the *Low* treatment and more optimistic in the *High* treatment).

The key finding of this paper is that students complete more lessons in Duolingo and are less likely to drop out of Duolingo in response to an increase in the perceived effectiveness of effort. In particular, using the instrumental variables strategy, I find that a one standard deviation (sd) increase in the perceived effectiveness of effort increases number of Duolingo lessons completed by 0.35 sd. Furthermore, a one sd increase in the perceived effectiveness of effort decreases the probability of dropping out, defined as not completing any lessons during the study period, by 10.4 percentage points.[6] Survey measures indicate that the findings are not driven by students shifting effort from one platform to another. Finally, although not statistically significant, students' final test scores also move in the same direction with the shifts in their beliefs. Overall, these results suggest that changing students' beliefs about the effectiveness of effort significantly influences their effort and that substitution effect dominates the income effect in this setting.

[6] Although I have effort data for all students in the baseline sample, belief data is missing for 5% of them. The results remain similar if I impute the minimum or maximum values for the missing observations.

These results have important implications for policy makers and education practitioners. Instructors and policy makers are often concerned about low levels of effort and try to identify low-cost ways of increasing effort. Given that information treatments are effective at influencing students' effort, a cost-effective way of increasing effort will be providing information regarding the average effort-performance relationship of previous participants of a course in settings where returns to effort are underestimated. Considering the importance of beliefs about returns to effort in effort decisions, existing educational surveys should incorporate more questions about beliefs about returns to effort as well as collect more data on students' effort decisions to further our understanding on the topic.

Will the results from this study generalize to traditional learning settings? Although a clear cut answer to this question will require more studies in these settings, thinking about the similarities and differences between remote and traditional learning environments is valuable. In both settings, effort has a cost and benefit associated with it, students can affect their learning by changing their effort, and students can choose leisure activities or studying for some other subject as an outside option. However, students are less experienced with Duolingo environment compared to a traditional learning setting which can be one reason for the high inaccuracy of beliefs about returns to effort. Finally, students voluntarily sign up for Duolingo whereas most traditional learning platforms are compulsory. The voluntary nature of the platform can be a reason of why students, on average, overestimate returns to effort. Hence, although beliefs about returns to effort may be different in different settings, these beliefs are likely to play a role in effort choices in all educational settings. This role of information provision on effort is expected to be small in settings where students get a lot of feedback about how effort translates into performance through other means, but large in settings where the feedback is scarce and noisy and where students are inexperienced.

This paper most closely relates to the growing literature on the role of beliefs in educational settings. While this literature documents the effect of beliefs about *returns to education* on students' *schooling* decisions (Nguyen, 2008; Jensen, 2010; Abramitzky and Lavy, 2014; Fryer, 2017) and *major* choices (Wiswall and Zafar, 2015; Hastings, Neilson and Zimmerman, 2015; Conlon, 2019)[7], this paper is the first to analyze the effect of beliefs about *returns to effort* on students' *effort*

[7]There is also a growing literature that studies how beliefs of parents about the human capital production function correlate with or affect their investments in education (Attanasio, Cattan, Fitzsimons, Meghir and Rubio-Codina, 2020; Cunha, Elo and Culhane, 2013; Boneva and Rauh, 2018; Dizon-Ross, 2019).

decisions. To my knowledge, this study is also the first to show that beliefs about *returns to effort* can be manipulated through information provision and that both anecdotal and average information can be used to manipulate these beliefs.

This paper also adds to the literature on performance feedback. Existing research (Azmat and Iriberri, 2010; Bandiera, Larcinese and Rasul, 2015; Bobba and Frisancho, 2016; Azmat, Bagues, Cabrales and Iriberri, 2019; Dobrescu, Faravelli, Megalokonomou and Motta, 2019) show that providing students feedback about their relative performance or past performance affects their performance.[8] Furthermore, Stinebrickner and Stinebrickner (2012) and Stinebrickner and Stinebrickner (2014) simulate that 40-45% of drop-out decisions in the first two years of college can be attributed to students' learning about their performance. Hence, these studies show that obtaining performance information *specific to oneself* have effects on performance and dropout decisions. My paper contributes to this literature by presenting that information about the *general* effort-performance relationship affects study effort.

2 Conceptual Framework

In this section, I present a simple framework of effort choice that incorporates beliefs about the effectiveness of effort. The purpose of this framework is to illustrate that the impact of change in beliefs on effort is ambiguous from a theoretical perspective. This framework is an application of the canonical labor supply model.[9]

Consider a student who gets utility from achievement, A, and leisure, L. The student has a total amount of time, $\overline{T}$, which can be used either for studying, S, or for leisure. Assume that achievement is a linear function of studying.[10] The student wants to maximize his utility, $U(A, L)$,

[8] By utilizing natural experiments, Azmat and Iriberri (2010) find that provision of relative performance feedback to students increases students' grades and Bandiera, Larcinese and Rasul (2015) find that providing students with feedback on their past exam performance positively affects their future exam performance. Through field experiments, Bobba and Frisancho (2016) demonstrate that providing students feedback about their performance in a mock placement exam affects their academic track choices, Azmat, Bagues, Cabrales and Iriberri (2019) show that providing feedback to college students on their position in the grade distribution reduces their educational outcomes, and Dobrescu, Faravelli, Megalokonomou and Motta (2019) find that providing students relative performance feedback increases their grades.

[9] Economists tend to model study effort in a manner that is analogous to labor supply, where the rewards to study effort are like wages (Owen, 1995). A key distinction between their models and labor supply models is that the rewards to study effort are assumed to be fairly far in the future - through later earnings and so on - as opposed to contemporaneous wages (Rosen, 1977; Becker, 1993).

[10] In most learning environments, achievement is not only a function of how much students study but also a function of their aptitude. The role of ability in achievement has long been recognized in economic models both theoretically

which is increasing and concave in its arguments. Formally,

$$\begin{aligned} &\max_{A,L} \; U(A,L) \\ &s.t. \\ &L + S = \overline{T}, \text{ and} \\ &A = \beta S \end{aligned} \tag{1}$$

where β is the student's belief about the rate at which studying translates into achievement. Note that β is analogous to wage in a labor supply model and it is both the price and the opportunity cost of leisure. Rewriting the equations, we have the following maximization problem:

$$\begin{aligned} &\max_{A,L} \; U(A,L) \\ &s.t. \\ &A + \beta L = \beta \overline{T}. \end{aligned} \tag{2}$$

Then, the solution to this problem is given by $\beta MU_A = MU_L$. Hence, marginal utility of an extra unit of achievement should be equal to marginal utility of an extra hour of leisure divided by the cost of this extra hour in terms of units of achievement.

Let's consider a student who finds out that studying is less productive than he thinks. That is, he experiences a decrease in β. Will this student work more or less? If leisure is a normal good (as it is generally assumed theoretically and shown empirically), substitution and income effects will go in opposite directions. Since leisure is less costly now, the student will consume more leisure and will study less (substitution effect). However, since the student feels his time endowment is less valuable now, he will consume less leisure and will study more. Overall, if the substitution effect dominates, a decrease in β will lead to a decrease in effort. If the income effect dominates, a decrease in β will lead to an increase in effort. Hence, the effect is theoretically ambiguous. This paper manipulates student's belief about the rate at which studying translates into achievement (β) to empirically test which of these effects dominate.

and empirically (Becker, 1993; Black, Devereux and Salvanes, 2005; Cunha and Heckman, 2007). In this paper, I abstract from the role of ability in achievement by random assignment of individuals into the information treatments.

3 Experimental Design

3.1 Design Preview

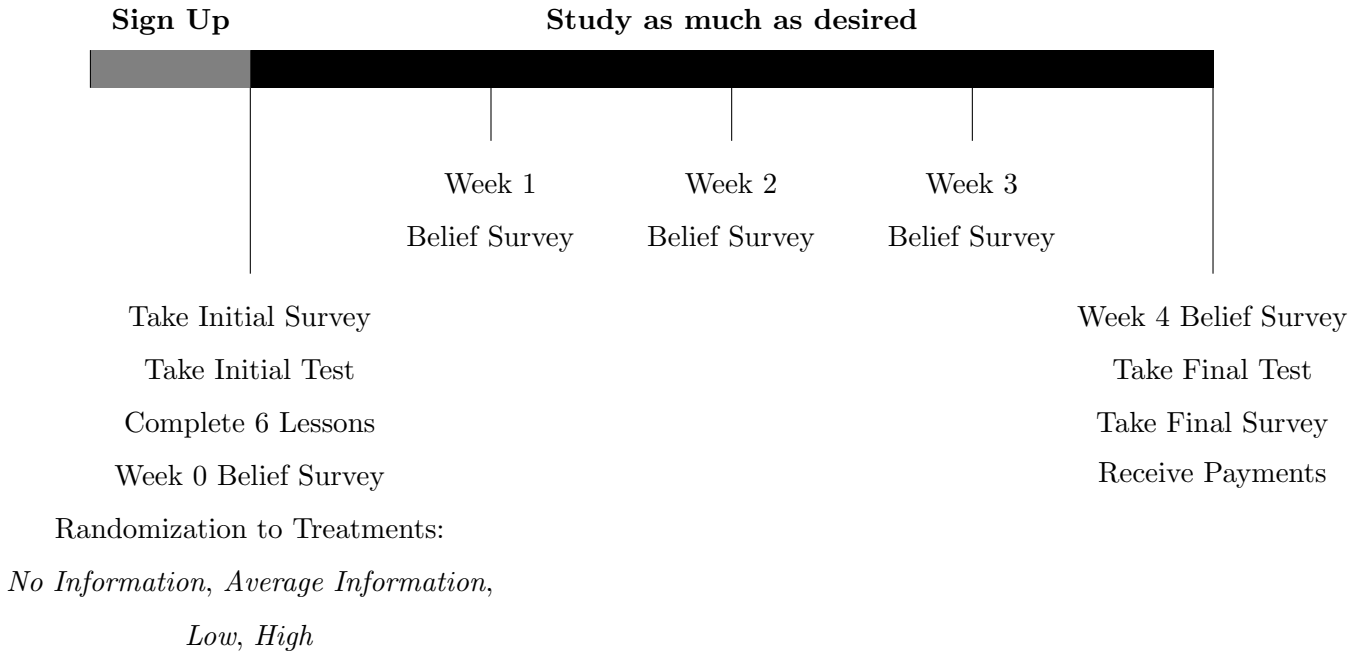
Figure 1 depicts the design of the experiment. From various Californian universities, intrinsically motivated individuals sign up to learn Spanish.[11] Eligible students complete a survey about their demographics, their Spanish background and certain personality traits. They take an online Spanish test which assesses their initial Spanish knowledge. Then, they sign up for the online language-learning platform, Duolingo, and complete 6 lessons to familiarize themselves with Duolingo. Every week, I elicit their beliefs in an incentivized way about how studying with Duolingo (number of Duolingo lessons completed) affects the test performance. After the first belief elicitation survey (Week 0 Belief Survey), students are randomly assigned to one of the four information treatments. In the *No Information* treatment, students do not receive any information about the relationship between effort and performance whereas in other treatments, they receive truthful information about this relationship. The information is based on the causal effort-performance relationship in Duolingo estimated for different participants from the same subject pool (Ersoy, 2020). In the *Average Information* treatment, students receive information about previous participants' improvement in test scores for completing 32, 64, and 96 lessons in a month. In the *Low* and *High* treatments, students receive information about the improvement of one of the previous participants, a low score and a high score, respectively. During the study period, students are free to study through the platform as much as they like for four weeks. After four weeks, they take another online Spanish test and answer a short survey. Students who successfully complete the study receive a completion payment of $40 and a bonus payment up to $30 based on the accuracy of their beliefs, both paid as Amazon e-gift cards at the end of the study.

3.2 Effort in Duolingo

Duolingo is a free popular platform for learning languages which has over 300 million users worldwide. Duolingo's approach is based on learning-by-doing. Duolingo has tasks that are associated with 4 different dimensions of language learning: reading, listening, speaking, and writing. Learners do tasks such as translating a sentence, answering multiple choice questions, typing the sentences they hear and repeating the sentences they hear.

[11] See Appendix A.1 for recruitment details. I choose Spanish because Spanish is the best developed course within Duolingo and the most popular language studied on Duolingo. Furthermore, preliminary surveys with students from the universities I recruit suggest that they would like to learn more Spanish.

Figure 1: Experimental Design



I conduct this study with Duolingo since it provides high internal validity due to observability, measurability, and homogeneity of effort. My main effort measure is lessons completed within Duolingo. Duolingo lessons, unlike traditional lessons, are a compilation of practice tasks (See Appendix Figure A.1 for examples of practice tasks in a beginner lesson). Students can make many mistakes in a lesson but need to answer all the tasks correctly to be able to complete a lesson. Furthermore, learners need to complete all the previous lessons to be able to unlock the upcoming lessons. Hence, Duolingo facilitates a high degree of standardization on outputs, in the sense of completed lessons. Since a task-standardized effort measure is adequate for the purposes of this study, I use administrative data on the number of new lessons completed in Duolingo as the main effort measure.

As an alternative effort measure, I use minutes spent on Duolingo. Since how much time one spent on a lesson can be affected by one's language learning ability, focus, and demographic background, minutes spent on Duolingo is not my preferred measure of effort. Additionally, I do not have administrative data on time spent on each lesson. I only have the lesson timestamps that indicate when a lesson was completed but do not provide information about when a lesson was started. To create minutes spent measure by using timestamps, I proxy the start time of a lesson with the end time of the previous lesson and calculate the time spent for each lesson. If the difference between start times of two successive lessons is more than 10 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.

3.3 Spanish Assessment Test

Students' Spanish knowledge is assessed with an online language test at the beginning and at the end of the study period. The initial and final tests are based on the practice tasks in Duolingo lessons. Each test consists of 80 questions. 40 of these questions are based on Duolingo Beginner Skills lessons and the other 40 of them are based on Intermediate Skills.[12] A person who has perfectly mastered Duolingo Beginner skills, Duolingo Intermediate skills (the next 142 lessons),

[12]Duolingo has 317 lessons in total of which 67 belong to Beginner Skills, 142 belong to Intermediate Skills and 108 belong to Advanced Skills. Since the targeted subject pool of this experiment is individuals with no or little knowledge of Spanish, I construct the test based on only the lessons in Beginner and Intermediate Skills. See Appendix Figure A.2 for the list of lessons during the study period.

and nothing else should be able to correctly answer all 80 questions of the test and get a score of 1000 points.[13]

Students are not paid for their performance in the test since I am interested in how changes in beliefs about the returns to effort affect effort in a setting where the monetary returns to the performance are far in the future. The reaction to information in such a setting can be quite different than the reaction to information in a setting of immediate monetary returns to performance. One solution would be to announce that the test is incentivized only right before the final test. But, such a solution could encourage cheating in the test. Even without monetary incentives, cheating can be an issue if the students get some instrumental value from scoring high on the exam no matter how that high score is achieved. Hence, I take two precautions to minimize cheating. First, I remind participants that they should not use any other sources while taking the exam as this is important for the validity of the research results. Second, I ask them to type "I agree that I will answer the test questions with my existing knowledge and I will not consult any other sources." right before the test.

3.4 Information

To be able to provide students factual and causal information about the returns to effort in Duolingo, I use data from a previous experiment (Ersoy, 2020). Students in both experiments are recruited from the same pool. In the previous experiment, I manipulate effort by randomly assigning participants into one of the five different lesson groups (8 lessons, 12 lessons, 16 lessons, 20 lessons, and 24 lessons per week for 4 weeks). Compliance with lessons assignments were high, creating the exogenous variation in effort that is needed to estimate the causal returns to effort. Using lessons assignments as an instrument for the number of lessons completed, I estimate how the number of lessons completed within Duolingo affects the test performance.

I randomly assign students in this experiment into one of the four treatments. In the *No Information* treatment, I do not provide any information about how the number of lessons completed within Duolingo affects the test performance. The *Average* treatment provides information on how much the participants in the previous experiment improve, on average, in terms of test scores after completing 32, 64, and 96 Duolingo lessons in a month (See Appendix Figure A.3a). The other two

[13] See Appendix A.2 for the test details.

treatments instead provide information about one of the participants of the previous experiment (anecdotal information). I choose two participants from the previous experiment such that both are assigned to complete the same amount of lessons and both complete those lessons but one improves little and the other improves a lot. Students in the *Low* treatment receive information about the participant who improves little whereas students in the *High* treatment receive information about the participant who improves a lot (See Appendix Figure A.3b and A.3c, respectively).[14]

Ex-ante, it is hard to tell which treatment intervention would work the best to manipulate the beliefs of the students regarding the effort-performance relationship. *Low* and *High* information treatments provide the opportunity for the cleanest comparison between groups. The students in both treatments receive the exact same information about how many lessons a previous participant completed. The only piece of information that differs between the two treatments is how much that previous participant improved in test scores. These treatments also allow me to randomly send negative news to some set of the students (information provided in the *Low* treatment is lower than what most students believe about the effectiveness of effort) and positive news to the others (information provided in the *High* treatment is higher than what most students believe about the effectiveness of effort) regardless of their initial beliefs. Furthermore, information in the *Low* and *High* treatments is less abstract since it is explicitly about a single participant, a participant the students can relate with in terms of initial test scores. However, since the information provided in these treatments is only about one participant, it might be unconvincing to change beliefs. To alleviate this concern, the *Average* treatment provides information about the average improvements in test scores for three different levels of effort.

There are many ways and means of providing information. I provide the information in a storyboard format since a storyboard is not only engaging but also easily reproducible.[15] In addition to providing information right after I elicit students' beliefs, I send them an email containing the information. This method of providing information is easily applicable in other online learning settings.

[14]The exact information the students obtain in the *Low* and *High* treatments varies slightly based on their initial test scores since I want to provide them information about another participant who is similar to them in terms of initial test scores. To do so, I divide the participants of the previous experiment into quartiles based on their initial test scores and match them with participants of this experiment based on these quartiles.

[15]See Appendix Figure A.4 for the storyboard of the *Average* treatment. The storyboard for other information treatments are the same except the last two figures. The storyboard for the *No Information* treatment consists of only the first four figures. All storyboards are created with `https://www.storyboardthat.com` and available upon request.

3.5 Belief Elicitation Surveys

Belief elicitation surveys aim to measure the beliefs of the students about how Duolingo effort affects performance in the test. Students take these surveys five times during the study: at the beginning of the study (Week 0), at the end of Week 1, at the end of Week 2, at the end of Week 3, and at the end of the study (Week 4). Students do not receive any feedback about the accuracy of their beliefs until the end of the experiment (except the information they receive from the information treatments). I choose to elicit their beliefs weekly because I want to measure how students update their beliefs as the study progresses. Frequent measurement of beliefs is important because it shows us whether inaccuracies or biases in beliefs are temporary or not.

In belief elicitation surveys, I ask students two types of belief questions: questions about the participants of the previous experiment (henceforth, other participants) and questions about themselves. Both questions are asked for five different levels of effort. The belief question about the other participants reads as follows: "On average, how much improvement in test scores do you think results from completing X lessons per week for 4 weeks (X*4 lessons in total)?" where $X \in \{8, 12, 16, 20, 24\}$. Before answering this question, all students know the average initial score of the other participants (see Appendix Figure A.5 for a screenshot). After students provide their answers for all five levels of effort, they see all of their answers on a single page and either confirm their answers or revise them until they are satisfied with their answers. Then, students answer belief questions about themselves. The belief question about the self reads as follows: "What do you think your improvement in test scores would be if you complete X lessons per week for 4 weeks (X*4 lessons in total)?" where $X \in \{8, 12, 16, 20, 24\}$. For this question, students are told to imagine that they were a participant in the previous experiment and their score in the initial test were equal to the average initial score of the previous participants (see Appendix Figure A.6 for a screenshot). The purpose of the belief questions about the self is to know how different the students think they are from other participants in terms of returns to effort. Since beliefs about the self cannot be measured in an incentivized manner and since beliefs about returns to effort for self can be confounded with beliefs about returns to ability, I use beliefs about the other participants as the main belief measure.[16]

[16] Paying subjects for the accuracy of their beliefs about the self would not be incentive-compatible since participants have the power of affecting their own test scores.

Incentivized belief elicitation is a widespread procedure that is used in many experiments to make sure that the participants pay attention to the questions and provide truthful answers.[17] Incentivized belief elicitation techniques are also used for measuring subjective probabilities in the field (Burks, Carpenter, Goette and Rustichini, 2013; Drerup, Enke and von Gaudecker, 2014; Augenblick, Cunha, Dal Bo and Rao, 2016). I pay students based on the accuracy of their answers in one of the five belief questions about the other participants in one of the five weekly surveys (randomly chosen) using an incentive-compatible mechanism.[18] In particular, I use an absolute deviation rule. For each point that students' guesses differ from the estimated average improvement of the other participants, 50 cents is discounted from their bonus payment of $30.

To make sure that students provide their beliefs in an informed manner, I supply all students with three sets of information before they take Week 0 Belief Survey. First, I provide them information about Duolingo lessons (see Appendix A.3). At that point, students are familiar with Duolingo lessons since they have already joined the platform and completed six Duolingo lessons. Second, I provide them detailed information on how the test is scored and what the test scores mean (see Appendix A.4). At that point, students have already taken the initial Spanish test. To be able to provide the students with information on what the test scores mean, I benchmark what the average test score of a student who has just finished the first semester of a college level Spanish course should be.[19] I further provide students information on what different test scores would mean in terms of being able to do certain activities (e.g. someone with a test score of 161 can order and pay at a cafe, etc). Third, I provide students some information about the previous experiment (see Appendix A.5). This includes a description of how the previous experiment is conducted and the average initial test score of the previous participants. I supply all students with sufficient information so that they are familiar with Duolingo lessons and they can understand what the test scores mean in terms of both levels of academic achievement and functional capability.

[17] For a literature review on belief elicitation techniques, see Schotter and Trevino (2014). In laboratory settings, Wang (2011) and Gächter and Renner (2010) find that beliefs are more accurate when they are incentivized.

[18] Week 0 survey is chosen for payment with 80% chance and each of the surveys in the other weeks has 5% chance of counting. I assign small weights to the belief questions in the other weeks since paying students based on accuracy of their beliefs might lead them to pay attention to the information treatments for spurious reasons, may increase the likelihood of contamination across treatment groups, and can cause anticipated income effects.

[19] For this benchmarking exercise, I recruit students from Stanford, UC San Diego and UC Irvine who are enrolled in a college level Spanish course and I give them my Spanish tests. By using their scores in the test and their level in their courses, I benchmark what the average test score of a student who has just finished the first semester of a college level Spanish course should be. I thank Professor Grant Goodall from UCSD and Professor Julio Torres from UCI for their help. See Appendix A.6 for the details of the benchmarking exercise.

3.6 Initial Survey

Since personality traits are important predictors of academic achievement, I assess participants' growth mindset (Paunesku, Walton, Romero, Smith, Yeager and Dweck (2015)), locus of control (Rotter (1966)), self control (Tangney, Baumeister and Boone (2004)), and big five personality traits (Rammstedt and John (2007)) in the first part of the initial survey. The Growth Mindset Survey consists of 8 statements with a 6-item Likert scale (answer choices range from "Disagree a Lot" to "Agree a Lot"). It asks students' beliefs and goals regarding ability and performance with statements such as "No matter how much intelligence you have, you can always change it a good deal.". The Locus of Control Survey asks questions about students' beliefs on how luck affects certain outcomes. It consists of 8 statement pairs such as "The idea that teachers are unfair to students is nonsense." vs "Most students don't realize the extent to which their grades are influenced by accidental happenings." Students need to choose the statement they believe more strongly from each pair. The Self Control Survey contains 13 statements which students rate based on how much the statements apply to them (on a 5-item Likert scale, ranges from "Not at all" to "Very Much"). The survey measures how much self control students have with statements like "I am good at resisting temptation." The Big Five Personality Traits Survey consists of 10 statements with a 5-item Likert scale (answer choices range from "Strongly Disagree" to "Strongly Agree"). Each 2 questions measures a separate personality trait. Personality traits measured are extroversion, agreeableness, conscientiousness, neuroticism, and openness.

In the second part of the initial survey, students answer various questions about their language background, their schooling background, and their demographics. In terms of the language background, I ask students their current level of Spanish knowledge, their purpose to learn Spanish, and their native language. In terms of the schooling background, I ask students their major, their class year, their college GPA, and their standardized test scores (e.g. SAT or ACT). At the end of the survey, I ask students their gender, ethnicity, highest level of education completed by their father and mother, monthly expenditure, and annual family income.

3.7 Preference Survey

After students complete the initial survey and initial Spanish assessment test, they take an incentivized survey that aims to elicit their time preferences (including time inconsistency) and risk preferences at an individual level. To elicit time and risk preferences of the students, I use the

Convex Time Budget (CTB) method (Andreoni and Sprenger, 2012).[20] After finishing this survey, students receive a link to sign up for Duolingo.[21]

3.8 Final Survey

At the end of the experiment, students take the final survey. In the final survey, I ask students some questions that would allow me to check for spillovers and potential confounds. I ask students whether they have used any online or offline tools to learn Spanish other than Duolingo during the study period. Students are also asked whether they have used any external resources (i.e. cheating) while taking the initial and final Spanish assessment tests, whether they have used another Duolingo account during the study period, whether they have participated in the previous rounds of the study. To check for potential contamination among friends, I ask students whether they have friends who have participated in the study and the name of the friend(s), if any. I also ask students some true/false questions to check whether they were aware of the details and the rules of the study.

Students in the *Low* and *High* treatments are also asked about who they think the information they were given belongs to. In particular, I remind the students that they were given information about a single participant from the previous experiment who had a similar initial test score to them. Then, to shed light on the mechanisms of why anecdotal information affect perceptions, I ask the students who they think this participant is in a multiple choice question with answer choices "The participant who improved the most", "The participant who improved the least", "A randomly chosen participant", "The participant who experienced median improvement", "The participant who experienced average improvement", and "I don't know".

4 Data

4.1 Summary Statistics and Balance

Appendix Table B.1 presents the summary statistics. Column (1) presents the mean values of the 36 control variables that were collected in the initial survey for 628 students who completed the initial survey (initial sample). 69% of the respondents are female. 40% of the respondents are

[20] I follow experimental procedures in Andreoni, Kuhn and Sprenger (2015). See Appendix A.7 for details.

[21] Students who do not want to take this survey are allowed to continue to sign up for Duolingo since this survey is not the main part of the experiment. Forcing students to complete this survey could result in them dropping out from the study all together hence I make this survey voluntary.

Asian and 34% are Caucasian. 66% of the survey takers report English as their native language. 59% of the respondents are enrolled in Health and STEM majors. 92% of them report that they do not know any Spanish or they know Spanish at the beginner level.

The experiment took place in two waves. The recruitment method and the experimental procedures were the same between the two waves. 284 students completed the initial survey in Wave 1 (March 5-April 7, 2017) and 344 students completed the initial survey in Wave 2 (June 24-July 26, 2017) across four treatment arms.[22] Appendix Table B.1 Columns (2) and (3) present the averages for Wave 1 and Wave 2, respectively. By regressing each control variable on a dummy variable for being a participant in Wave 2, I test the balance across the two waves. Appendix Table B.1 Column (9) reveals that there are many statistically significant differences across the two waves. However, the differences in observables across the waves are unlikely to confound the treatment effects since the identifying variation comes from the treatments, not from the waves. Nevertheless, I control for the wave fixed effects in the analysis.

Appendix Table B.1 Column (4) presents the mean values of the 41 control variables for 391 students who completed the initial survey, who took the initial test, who joined Duolingo and who took the baseline belief survey (baseline sample).[23] Comparing Columns (1) and (4), we see that there are some differences in observables across the initial sample and baseline sample. P-values are reported in Column (10) and obtained from regressing each control variable on a dummy for being in the baseline sample. Compared to the students in the initial sample, students in the baseline sample are more likely to be Asians, have higher GPAs and standardized SAT/ACT math scores, have more educated fathers, are more likely to be sophomores and less likely to be seniors, are more likely to major in Health and STEM and less likely to major in Business and Economics, and have more self-control.

[22] In Wave 1, there were two additional treatments: the *32-Lesson* treatment and the *96-Lesson* treatment. The *32-Lesson* (*96-Lesson*) treatment provided information about the average improvement in test scores only for completing 32 (96) lessons. Since both the number of lessons completed and the effectiveness of effort differ across these two treatments, why these treatments affect effort would not be clear if these treatments affected students' effort. Hence, these treatments were excluded from the design of the second wave and their data from Wave 1 are excluded from the analysis.

[23] I drop one student who reported that he/she participated in the previous experiment. I drop two students with the same email addresses. I drop one extreme outlier for whom the number of lessons completed was 11 sd away from the mean and 7 sd away from the second largest observation.

To test balance across treatment arms, Appendix Table B.1 Columns (5) - (8) present the mean values of the control variables for each treatment and Column (11) reports the F-statistics for the coefficients across treatments being equal to each other from the regressions of each control variable on treatment dummies. I find that we can reject the joint null hypothesis of equality for 5 out of 41 of these variables (Caucasian, Freshman, Growth Mindset, Openness, Native English Speaker). Hence, I present the findings with and without an extensive set of controls to check the robustness of the results.

4.2 Attrition

95% of the students in the baseline sample (371 out of 391) complete Week 1 belief survey and 83% of the students in the baseline sample (324 out of 391) complete the study by taking the final test and final survey. The dependent variable (effort) is available for all the students in the baseline sample. The endogenous variable of interest (change in students' beliefs from baseline to Week 1) is only available for the students who complete Week 1 belief survey. Hence, my analysis sample consists of 371 students who complete Week 1 belief survey. Appendix Table B.2 Column (2) presents the summary statistics of the analysis sample whereas Column (1) replicates the statistics of the baseline sample for easy comparison. P-values reported in Column (7) are obtained from regressing each control variable on a dummy for being in the baseline sample. Compared to the students in the baseline sample, students in the analysis sample are more likely to be females, have higher openness and neuroticism scores as well as higher GPAS, less likely to be seniors, African-Americans, or risk-averse, and less likely to miss the preference survey.

To test differential attrition, Appendix Table B.2 Columns (3) - (6) present the averages of the control variables in each treatment group for the analysis sample. Column (8) reports the F-statistics for the coefficients across treatments being equal to each other from the regressions of each control variable on treatment dummies. I find that we can reject the joint null hypothesis of equality for 7 out of 41 of these variables (Asian, Caucasian, Freshman, Growth Mindset, Openness, Native English Speaker, Patience). However, since almost all of these variables are unbalanced at the baseline sample, it is hard to know whether there is differential attrition based on treatment assignments. Nevertheless, the results of the paper remain similar if I use the baseline sample and impute the minimum or the maximum values for the missing observations of the change in beliefs in the analysis sample.

5 Findings

5.1 Baseline Beliefs

Figure 2 Panels (a)-(e) show the cumulative distributions of the baseline beliefs about the improvement of other participants for five different levels of effort (32, 48, 64, 80, and 96 lessons).[24] We see that there is substantial heterogeneity in the baseline beliefs. The solid lines depict the *estimated* average returns to Duolingo effort and the dashed lines depict the 95% confidence intervals, using the data in Ersoy (2020). As seen from the figure, subjects, on average, overestimate the returns to Duolingo for all five levels of effort and they overestimate more for higher levels of effort. If I classify subjects based on their beliefs, I find that 53% of the subjects overestimate the returns to Duolingo in all five questions and 35% of the subjects underestimate the returns to Duolingo in all five questions. The rest of the subjects underestimate the returns in some questions and overestimate them in others.

To study the baseline beliefs in a concise manner, I define a composite measure which averages the answers across all five belief questions and calculates the perceived per-lesson improvement in the following manner:

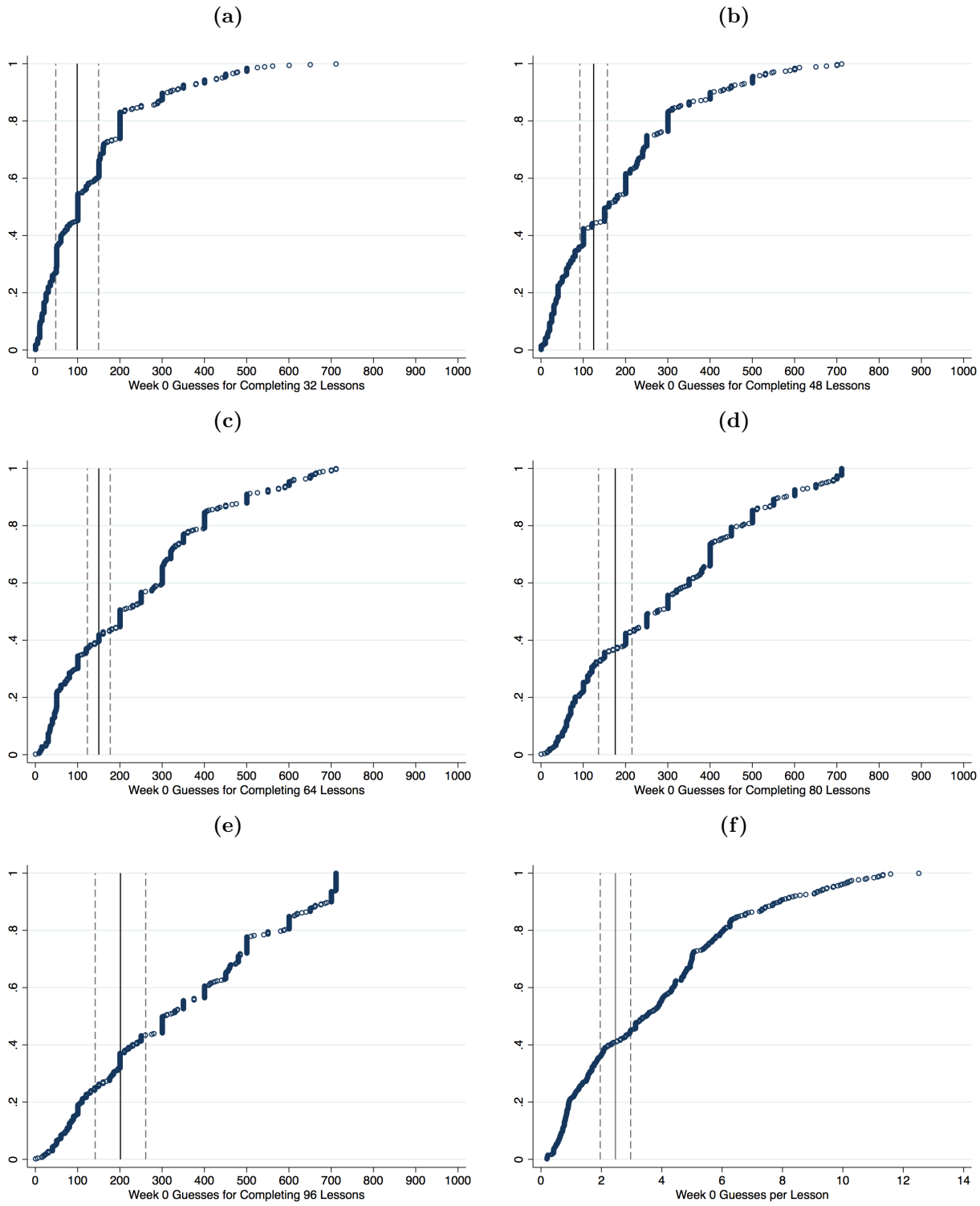
$$Baseline\ Belief\ per\ Lesson = \frac{1}{5} \sum_{X \in \{32,48,64,80,96\}} \frac{Week\ 0\ Belief_X}{X}. \tag{3}$$

where $Week\ 0\ Belief_X$ is a subject's answer to the belief question for X lessons where $X \in \{32, 48, 64, 80, 96\}$ in the Week 0 Belief Elicitation survey. According to this measure, subjects believe that the average improvement in test scores per lesson is 3.81 with a standard deviation of 2.83. The median baseline belief per lesson is 3.49, the 25th percentile is 1.25, and the the 75th percentile is 5.54. Figure 2 Panel (f) shows the cumulative distribution of this composite belief measure. The *estimated* average per-lesson improvement is 2.47, depicted with the solid line (95% confidence interval is depicted with the dashed lines). Hence, the perceived average per-lesson improvement is approximately 50% larger than the estimated average and 59% of the subjects overestimate returns to effort according to this measure.[25]

[24] Since the highest possible belief one can hold for the improvement is 711 points (maximum possible test score (1000 points) minus the average initial score of other participants (289 points)), I top coded the beliefs that are higher than 711 points as 711 points throughout the analysis in the paper. Approximately 10% of the beliefs fall into this category.

[25] Appendix Table B.3 shows how baseline beliefs correlate with subjects' observable characteristics. African-Americans, subjects with a growth mindset, and more risk-averse individuals hold more optimistic beliefs about the

Figure 2: Cumulative Distribution Function of Beliefs about the Improvements of the Other Participants



Notes: Panels (a)-(e) depict subjects' answers to the belief questions about the other participants in Week 0 belief survey. In particular, panel (a) depicts subjects' answers to the question of how much improvement in test scores results from completing 32 lessons in a month, on average. Similarly, panels (b)-(e) depicts the answers for completing 48 lessons in a month, 64 lessons in a month, 80 lessons in a month, and 96 lessons in a month, respectively. See Appendix Figure A.5 for a screenshot of the belief questions. Beliefs are top coded at 711 points, which is the maximum possible test score minus the average initial score of other participants. Panel (f) depicts baseline beliefs for per-lesson improvement as defined in Equation 3. The solid lines depict the *estimated* average returns to Duolingo effort and the dashed lines depict the 95% confidence intervals, using the data in Ersoy (2020).

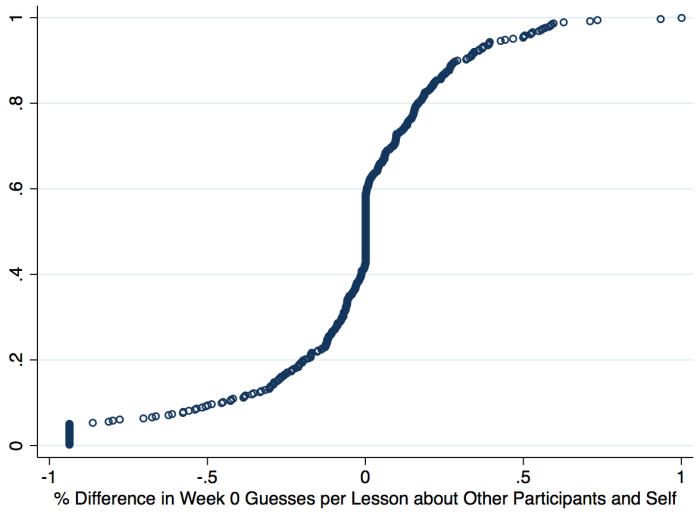
To analyze how beliefs about the self and the other participants differ, I first look at the correlation between the two composite belief measures. I find that they are highly correlated with a correlation coefficient of .84 (statistically significant at the 1% level). Then, I calculate the difference between the perceived per-lesson improvement for the others and for the self. I find subjects do not systematically think that they are more or less capable of learning than the other participants. The average difference is -.042 with a standard deviation of 1.60 and the median difference is 0.[26] Indeed, there are 64 subjects (16% of all subjects who took Week 0 belief survey) for whom there is no difference between the baseline beliefs about the self and about the others. This finding is striking given that subjects answer these two sets of belief questions on different pages of the survey, that they cannot go back and forth between the pages, and that they were not aware that they need to answer belief questions about the self when they answer belief questions about the other participants. Figure 3 displays the cumulative distribution of the percentage difference between the perceived per-lesson improvement for the others and the self. This distribution is quite concentrated. The first quartile of the distribution corresponds to a -12% difference and the third quartile corresponds to a 13% difference. Instead, when I randomly match initial beliefs about the self and other participants five hundred times, I observe the first quartile of these distributions ranges between -110% and -190% and the third quartile ranges between 48% and 63%. Hence, beliefs about per-lesson improvement for the others and self are quite similar.

Why do participants overestimate the returns to effort? One potential explanation is that participants are overconfident about their ability to learn Spanish. However, this explanation is unlikely since the participants overestimate the returns to effort not only for themselves but also for others. Another potential explanation is self-selection into the research study, similar to the winner's curse (Thaler, 1988). Participants who believe that studying with an online language learning platform is effective would select into the research study whereas the others would not. This selection mechanism would create an upward bias in beliefs. Similar selection mechanisms are likely to be existent for most voluntary learning environments. This finding might also be explained by a tendency to allocate one's attention in motivated ways (Bénabou, 2015). A growing body of evidence shows that beliefs are updated more in response to information that requires one to update his views in a

returns to Duolingo effort. Subjects with higher SAT/ACT reading scores and more present-biased individuals hold less optimistic beliefs.

[26] I analyze the difference in beliefs about the self and the other participants in a fixed effect regression framework, separately for each belief question and for the composite measure. Appendix Table B.4 shows that subjects' average beliefs about the self and the other participants are *not* statistically different from each other for any of these measures.

Figure 3: Cumulative Distribution Function of the Percentage Difference between the Perceived Per-Lesson Improvement of the Others and the Self



Notes: This figure depicts the percentage difference between the perceived per-lesson improvement of the other participants and the perceived per-lesson improvement for the self. -1 indicates that subjects believe that other participants' returns to effort is 100% worse than their returns to effort and 1 indicates that subjects believe that other participants' returns to effort is 100% better than their returns to effort. This distribution has a long left tail so the bottom 5% of the values are replaced with the value of the fifth percentile, -.94, for the purposes of the figure. All the statistics reported in the text and in the tables use the actual values.

positive direction than a negative direction[27], which might result in upwardly biased beliefs about the returns to effort.

5.2 Belief Updating

Within the first week of the experiment, subjects update their beliefs in line with the information they receive. Table 1 presents how beliefs about per-lesson improvement of the others change from Week 0 to Week 1. I control for baseline beliefs about the other participants for five levels of effort in each specification to account for regression to the mean. Several important patterns emerge. First, although subjects in the *No Information* treatment update their beliefs downwards (towards the *estimated* average returns to effort), the magnitude of this update is small and not statistically significant. Second, the information provided in the *Average* treatment is effective at decreasing the perceived effectiveness of effort (statistically significant at 1% level). Third, the anecdotal information provided in the *Low* treatment also decreases the perceived effectiveness of

[27]See Mobius, Niederle, Niehaus and Rosenblat (2011) and Eil and Rao (2011) for laboratory evidence and see Sharot and Garrett (2016) for a review.

effort (statistically significant at 1% level). The effects for the *Average* and *Low* treatments are not statistically significantly different from each other. Finally, I observe that anecdotal information provided in the *High* treatment increases the perceived effectiveness of effort (statistically significant at 10% level).[28] Hence, providing information about the averages or anecdotal information seems to be effective at changing beliefs. Including various sets of controls does not change the results.

After the first week of the experiment, subjects continue to update their beliefs in all treatments. Appendix Table B.5 presents how beliefs about per-lesson improvement of the others change from Week 1 to Week 4. Subjects in the *No Information* treatment continue to update their beliefs downwards (statistically significant at the 5% level) as they become more familiar with the learning environment. Subjects in the *Average* treatment update similar to the subjects in the *No Information* treatment. However, subjects in the *Low* and *High* treatments continue to update their beliefs downward (statistically significant at the 10% level) and upward (statistically significant at the 1% level), respectively, towards the information provided. The results remain similar as I include various sets of controls.[29]

Although the experiment in this paper is not designed to test the hypothesis of that providing anecdotal information can be as effective as providing information about the sample average to change beliefs, Table 1 and Appendix Table B.5 provide evidence supporting this hypothesis. This finding suggests that beliefs might be formed through anecdotes or from a few random observations and has an important policy implication. As an example, consider household energy consumption. Schultz, Nolan, Cialdini, Goldstein and Griskevicius (2007) find that households reduce their energy consumption after learning that their energy consumption is more than the average neighborhood usage and increase their energy consumption after learning that their energy consumption is less than the average neighborhood usage. Although the first effect is desirable from a policy perspective, the second is not. The findings from my experiment suggest that the policy maker does not need to reveal the average neighborhood usage to affect energy consumption and can affect it successfully by providing information about one neighbor with low energy consumption.

[28] Although the absolute magnitude of the updating in the *High* treatment is smaller than the one in the *Low* treatment, the effect sizes are comparable if I consider the percentage change given the information shock.

[29] To analyze whether subjects find the information provided relevant to themselves, I look at how beliefs about per-lesson improvement about the self change from Week 0 to Week 1 (Appendix Table B.6) and from Week 1 to Week 4 (Appendix Table B.7). Results show that subjects update their beliefs about returns to effort similarly regardless of them being about themselves or about others.

Table 1: Effect of Treatments on Perceived Returns to Effort of Other Participants: Week 0 to Week 1

	(1)	(2)	(3)	(4)	(5)	(6)
	Week 1 Belief per Lesson-Baseline Belief per Lesson					
Average	-0.925***	-0.933***	-0.956***	-0.969***	-0.936***	-0.982***
	(0.303)	(0.301)	(0.304)	(0.312)	(0.321)	(0.325)
Low	-1.169***	-1.154***	-1.168***	-1.215***	-1.199***	-1.229***
	(0.317)	(0.317)	(0.314)	(0.316)	(0.325)	(0.331)
High	0.648*	0.663**	0.637*	0.621*	0.704**	0.629*
	(0.338)	(0.337)	(0.339)	(0.344)	(0.352)	(0.359)
Constant	2.074***	2.692***	2.600**	4.101*	12.18	13.15
	(0.285)	(0.481)	(1.202)	(2.146)	(10.57)	(10.77)
No-Information:						
Mean	-0.357	-0.305	-0.303	-0.267	-0.282	-0.130
Standard Deviation	0.217	0.221	0.222	0.246	0.256	0.288
P-values:						
Average=Low	0.431	0.480	0.499	0.432	0.421	0.472
Average=High	0	0	0	0	0	0
Low=High	0	0	0	0	0	0
Controls:						
Baseline Beliefs	Yes	Yes	Yes	Yes	Yes	Yes
Wave Fixed Effect	No	Yes	Yes	Yes	Yes	Yes
Demographic	No	No	Yes	Yes	Yes	Yes
Education	No	No	No	Yes	Yes	Yes
Personality & Preference	No	No	No	No	Yes	Yes
Other	No	No	No	No	No	Yes
Observations	371	371	371	371	371	371

Notes: OLS Regressions. Dependent variable is the perceived per lesson improvement of the other participants elicited in Week 0 subtracted from the perceived per lesson improvement of the other participants elicited in Week 1. Average is a dummy variable that is equal to 1 if a student is in the *Average* treatment and 0 otherwise. Similarly, Low (High) is a dummy variable that is equal to 1 if a student is in the *Low* (*High*) treatment and 0 otherwise. The omitted category is the *No Information* treatment. Baseline Beliefs are subjects' beliefs about the improvement of other participants for 5 different levels of effort (32, 48, 64, 80, and 96 lessons) elicited in Week 0. See notes of Appendix Table B.1 for detailed description of the control variables. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

To provide evidence on why anecdotal information works well, in the final survey I ask subjects in the *Low* and *High* treatments whose information they think they received during the experiment. 55% of the subjects believe that the information belongs to the average or the median participant whereas 24% of them believe that the information belongs to a random participant. The remain-

ing participants state that they do not know (9%), they believe that the information belongs to the participant who improved the most (8%) or they believe that the information belongs to the participant who improved the least (4%). Appendix Table B.8 looks at whether the change in beliefs differ between subjects who believe that the information is based on the average/median participant or based on some other participant. Although the differences-in-differences coefficient is positive (showing that the updating is more prominent in the *High* treatment for the subjects who believe that information belongs to the average/median participant), it is far from statistical significance.

5.3 The Effects of Perceived Productivity on Effort

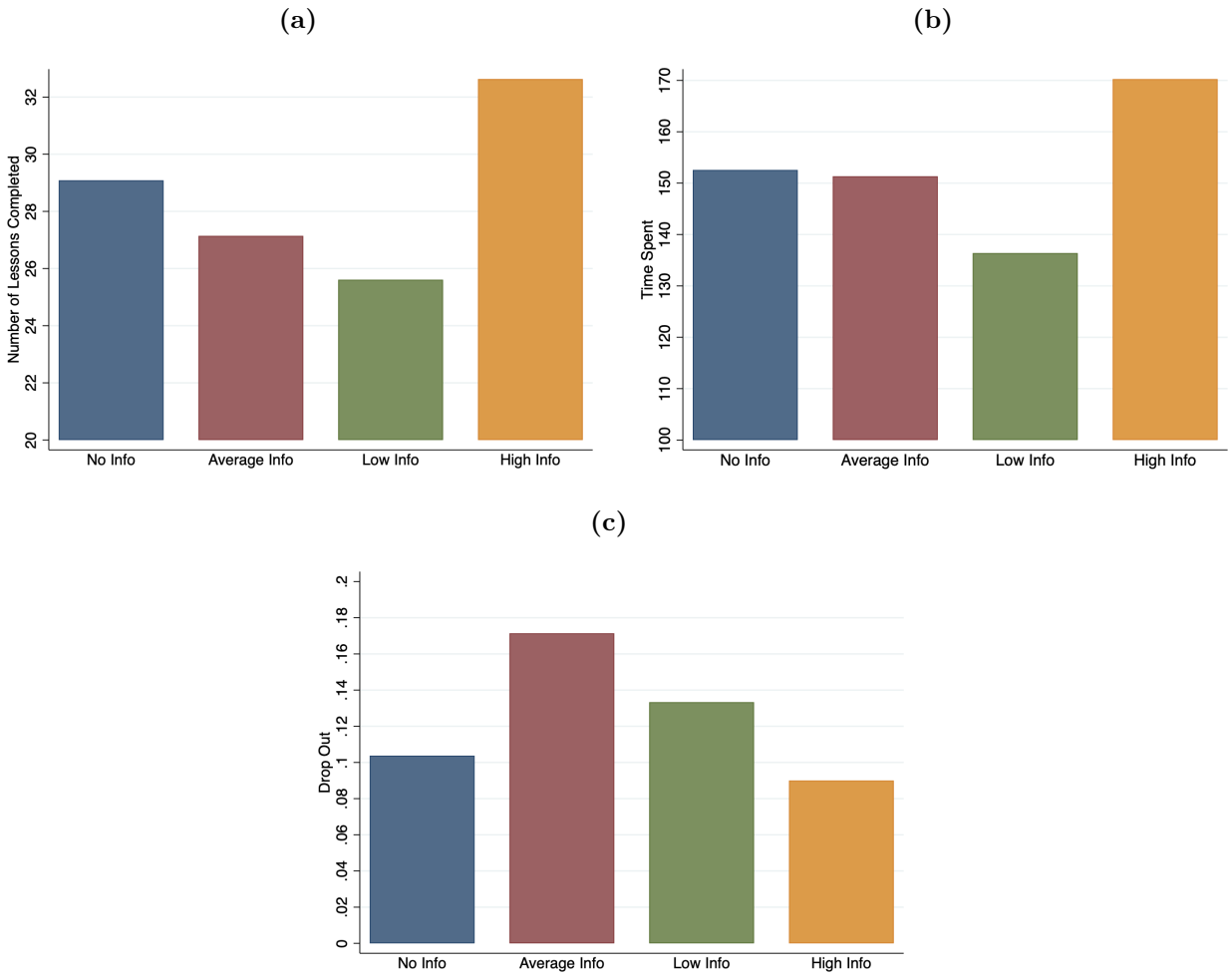
In this section, I analyze effort choices of the subjects in the intensive margin (number of lesson completed and time spent within Duolingo) and the extensive margin (dropping out). Figure 4 Panel (a) depicts the average number of Duolingo lessons completed, Panel (b) depicts the average time spent in Duolingo, and Panel (c) depicts the average dropout rate in each treatment group. On average, students complete 30.5 lessons and spend 152.5 minutes on Duolingo in the *No Information* treatment. 7 percent of the students in the *No Information* treatment drop out, meaning that they do not complete any Duolingo lessons after the first belief elicitation survey. Students complete fewer lessons in the *Average* and *Low* treatments and more lessons in the *High* treatment compared to the *No Information* treatment. Students spend less time on Duolingo in the *Low* treatment and more time on Duolingo in the *High* treatment. Students are more likely to drop out in the *Average* and *Low* treatments. These patterns, although not statistically significant in most cases, are in line with the changes in beliefs discussed in the previous subsection.

To study how changes in the beliefs about returns to effort *affect* effort, I use an instrumental variables (IV) approach and instrument the changes in beliefs from Week 0 to Week 1 with the treatment dummies. In particular, I run the following equations:

$$Change\ in\ Beliefs_i = \beta_0 + \sum_{K \in \{Average,\ Low,\ High\}} \beta_{1K} Treatment_{Ki} + C_i\Delta + \nu_i \tag{4}$$

$$Y_i = \alpha_0 + \alpha_1 Predicted\ Change\ in\ Beliefs_i + C_i\Gamma + \epsilon_i \tag{5}$$

Figure 4: Effort across Treatment Groups



Notes: Panel (a) depicts the average number of Duolingo lessons completed, Panel (b) depicts the average time spent in Duolingo, and Panel (c) depicts the average dropout rate in each treatment group.

where Y_i is the number of lessons completed within Duolingo, time spent within Duolingo, or whether a student drops out and C_i is the set of controls.[30] An IV approach is necessary to recover the causal effects since both the changes in beliefs and effort choices might correlate with unobservable characteristics of the students, the changes in beliefs might be measured with error leading to attenuation bias, and the amount of effort students put might affect students' beliefs about returns to effort (reverse causality).

Table 2 Panel A presents how the change in beliefs from Week 0 to Week 1 affects the number of lessons completed (α_1). Baseline beliefs are controlled in all specifications. According to Column (1), a one unit increase in the perceived effectiveness of effort increases the number of lessons completed by 4 (p-value: 0.054). Equivalently a one standard deviation increase in the the perceived effectiveness of effort (2.9 units) increases the number of lessons completed by 0.42 sd (1 sd in the number of lessons completed is equivalent to 27.8 lessons). When I control for a rich set of observables (Column (6)), the coefficient becomes 3.3 lessons (p-value: 0.074), which corresponds to a 0.35 sd increase in effort in response to a 1 sd increase in perceived productivity. This finding suggests that the substitution effect is larger than the income effect.

Table 2 Panel B presents how the change in beliefs from Week 0 to Week 1 affects time spent on Duolingo. An increase in the perceived effectiveness of effort increases time spent in Duolingo by 13 to 18 minutes (which corresponds to a 0.24 sd-0.32 sd increase in effort in response to a 1 sd increase in perceived productivity) depending on the specification, but none of these coefficients are statistically significant at the 10% level (p-values range between 0.123 and 0.238).

Table 2 Panel C presents how the change in beliefs from Week 0 to Week 1 affects the dropout behavior. I categorize a subject as dropped out if the subject does not complete any Duolingo lessons during the 4-week study period. According to Column (1), a one unit increase in the perceived effectiveness of effort decreases the probability of drop out by 3.6 percentage points (p-value: 0.078) from a mean of 12.5 percentage points which corresponds to a 29% decline. Equivalently a one standard deviation increase in the the perceived effectiveness of effort (2.9 units) decreases the probability of drop out by 10.4 percentage points. The coefficient becomes 4 percentage points (p-value: 0.042) when I control for a rich set of observables (Column (6)).

[30]Note that this specification is equivalent to the specification where belief in Week 1 is used as the endogenous variable (instead of the changes in beliefs from Week 0 to Week 1).

Table 2: Effect of Change in Beliefs on Effort

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A:	Number of Lessons Completed					
Change in Beliefs	4.003*	3.975*	3.398*	3.228*	3.385*	3.297*
	(2.074)	(2.073)	(1.974)	(1.902)	(1.879)	(1.844)
Constant	25.17***	21.15***	-22.54	-48.66**	-38.92	-20.70
	(4.711)	(6.996)	(14.64)	(22.67)	(103.8)	(104.4)
Panel B:	Time Spent on Duolingo					
Change in Beliefs	17.73	17.67	14.68	13.52	13.94	13.15
	(11.51)	(11.50)	(11.06)	(10.92)	(11.23)	(11.14)
Constant	127.6***	124.9***	-116.8	-202.7	-250.1	-235.8
	(25.33)	(41.89)	(95.34)	(153.9)	(529.1)	(546.2)
Panel C:	Drop Out (No Lessons Completed)					
Change in Beliefs	-0.0356*	-0.0358*	-0.0343*	-0.0322*	-0.0379**	-0.0396**
	(0.0202)	(0.0201)	(0.0195)	(0.0193)	(0.0192)	(0.0195)
Constant	0.167***	0.144**	0.337***	0.405*	0.950	1.056
	(0.0468)	(0.0734)	(0.123)	(0.220)	(1.401)	(1.352)
Controls:						
Baseline Beliefs	Yes	Yes	Yes	Yes	Yes	Yes
Wave Fixed Effect	No	Yes	Yes	Yes	Yes	Yes
Demographic	No	No	Yes	Yes	Yes	Yes
Education	No	No	No	Yes	Yes	Yes
Personality & Preference	No	No	No	No	Yes	Yes
Other	No	No	No	No	No	Yes
Observations	371	371	371	371	371	371
First Stage F-Stat	12.01	12.12	12.58	13.32	14.03	14.02

Notes: 2SLS Regressions. Dependent variable in Panel A is the number of lessons completed during the four weeks of the experiment. Dependent variable in Panel B is minutes spent on Duolingo during the four weeks of the experiment. Dependent variable in Panel C is a dummy variable which is equal to 1 if a subject does not complete any Duolingo lessons during the study period (hence drops out from Duolingo) or 0 otherwise. Change in beliefs is the perceived per lesson improvement of the other participants elicited in Week 0 subtracted from the perceived per lesson improvement of the other participants elicited in Week 1. The instruments are the treatment dummies. Baseline Beliefs are subjects' beliefs about the improvement of other participants for 5 different levels of effort (32, 48, 64, 80, and 96 lessons) elicited in Week 0. See notes of Appendix Table B.1 for detailed description of the control variables. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Change in beliefs about the effectiveness of Duolingo might affect effort choices outside of Duolingo. Although I am unable to observe participants' effort choices outside of Duolingo, I ask them in the final survey whether they have used any online or offline tools to learn Spanish other than Duolingo during the study period. 14% of the participants reported using other tools. Table 3 shows how changes in the beliefs affect self-reported usage of other tools to learn Spanish. According to Column (1), a one unit increase in the perceived effectiveness of effort decreases the usage of other tools by 4.1 percentage points) from a mean of 13.8 percentage points (p-value: 0.138). The effect remains similar in magnitude but is never statistically significant as I control for a rich set of observables (p-value: 0.182).

Table 3: Effect of Change in Beliefs on Self-Reported Effort outside of Duolingo

	(1)	(2)	(3)	(4)	(5)	(6)
	Use Any Other Tools to Learn Spanish					
Change in Beliefs	-0.0406	-0.0401	-0.0364	-0.0322	-0.0331	-0.0375
	(0.0273)	(0.0274)	(0.0273)	(0.0255)	(0.0278)	(0.0280)
Constant	0.232***	0.177	0.418**	0.236	2.463	2.177
	(0.0683)	(0.111)	(0.202)	(0.296)	(1.573)	(1.623)
Controls:						
Baseline Beliefs	Yes	Yes	Yes	Yes	Yes	Yes
Wave Fixed Effect	No	Yes	Yes	Yes	Yes	Yes
Demographic	No	No	Yes	Yes	Yes	Yes
Education	No	No	No	Yes	Yes	Yes
Personality & Preference	No	No	No	No	Yes	Yes
Other	No	No	No	No	No	Yes
Observations	324	324	324	324	324	324
First Stage F-Stat	11.90	11.89	12.53	13.49	13.11	13.03

Notes: 2SLS Regressions. Dependent variable is a dummy variable which is equal to 1 if a subject uses any online or offline tools to learn Spanish (other than Duolingo) during the study period or 0 otherwise. Change in beliefs is the perceived per lesson improvement of the other participants elicited in Week 0 subtracted from the perceived per lesson improvement of the other participants elicited in Week 1. The instruments are the treatment dummies. Baseline Beliefs are subjects' beliefs about the improvement of other participants for 5 different levels of effort (32, 48, 64, 80, and 96 lessons) elicited in Week 0. See notes of Appendix Table B.1 for detailed description of the control variables. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Finally, I explore whether the change in beliefs about returns to effort have an effect on Spanish test scores subject to the caveat that this experiment is not well powered to detect these effects. I

regress the standardized scores from the final assessment test on the standardized scores from the initial assessment test and the change in beliefs from Week 0 to Week 1, where the change in beliefs are instrumented with the treatment dummies. Appendix Table B.9 presents the results. The effects, although not statistically significant, are positive and consistent with the experimentally estimated effects in Ersoy (2020). Ersoy (2020) documents that 1 sd increase in the number of lessons completed in Duolingo increases the final test scores by 0.32 sd. In this paper, I find that 1 sd change in beliefs leads to 0.35 sd change in the number of lessons completed and 0.11 sd change in the final test scores. Hence, a back-of-the-envelope calculation shows that that 1 sd change in the number of lessons completed leads to 0.31 sd change in the final test scores according to estimates from this paper, which is consistent with the experimentally estimated effect.

5.3.1 Robustness Checks

I conduct robustness checks of the results by restricting the sample based on the responses in the final survey. Since not everyone in the analysis sample took the final survey, I first replicate Table 2 Column (6) for the final survey takers. As Appendix Table B.10 Column (1) shows, the results are more statistically significant for the final survey takers. In particular, a one unit increase in the perceived effectiveness of effort increases the number of Duolingo lessons completed by 4.4 (p-value: 0.031) and decreases the probability of drop out by 5.6 percentage points (p-value: 0.004). Then, Appendix Table B.10 Columns (2)-(6) present the main results for the subsample of individuals who did not use any online or offline tools to learn Spanish (other than Duolingo), who did not use any sources while taking the initial and final Spanish tests, who did not use another Duolingo account during the study period, who did not have friends participated in the study, and who correctly answered at least three questions (out of four questions) about the rules of the study. The results continue to hold for these subsamples.

I also conduct a bounding exercise to deal with missing observations in the analysis sample. Although 391 students are in the baseline sample (98 in *No Information*, 99 in *Average Information*, 98 in *Low*, and 96 in *High* treatments, respectively), only 371 students take the Week 1 survey (92 in *No Information*, 94 in *Average Information*, 93 in *Low*, and 92 in *High* treatments, respectively). Hence, the change in beliefs variable is missing for 20 students even though effort measures are recorded for all of the students. To check the robustness of the results to the inclusion of these missing observations, I impute the minimum or the maximum values of the change in beliefs variable

for the missing values. The results are shown in Table B.11 and Table B.12, respectively. The estimates barely change as a result of this bounding exercise.

Finally, Table 2 uses the change in beliefs and baseline beliefs about the others since the beliefs about the others are elicited in an incentivized manner and are not confounded by beliefs about ability. Appendix Table B.13 replicates Table 2 using the change in beliefs and baseline beliefs about the self. The results become noisier but remain similar in magnitude.

5.3.2 Heterogeneity by Personality Traits

Personality traits are important predictors of academic achievement.[31] The existing literature documents that locus of control is strongly correlated with academic achievement (Coleman, Campbell and Hobson, 1966; Findley and Cooper, 1983) and labor market outcomes (Cobb-Clark, 2015).[32] Alan, Boneva and Ertac (2019) show that grit, a related concept to the locus of control, is malleable through education interventions and affect students' effort behavior and academic achievement. Relatedly, Duckworth (2016) argues having grit is a key determinant of achievement. Dweck (2006) argues holding a growth mindset, believing that ability is malleable and can be changed through effort, is an important determinant of achievement.

In this setting, we might expect that the same information might lead to different types of behavior depending on one's personality traits. Suppose that a student updates his beliefs about the effectiveness of Duolingo downwards. This downward shift in beliefs would not necessarily translate into lower effort for every student. Students with a growth mindset might increase their effort to do better whereas participants with a fixed mindset might give up trying. Conversely, suppose that a student updates his beliefs upwards. This upward shift in beliefs would not necessarily translate into higher effort for everyone. Students with high self control or high locus of control (those who believe that their outcomes are mostly due to internal factors) might be able to put more effort to reap the benefits whereas students with low self control or low locus of control might not be able to do so.

[31] See Almlund, Duckworth, Heckman and Kautz (2011) and Heckman and Kautz (2014) for a review on the importance of various personality traits in various domains.

[32] Locus of control is one of the most frequently studied concepts in psychology (Rotter, 1990; Lefcourt, 1992) and it is included in many educational and labor force surveys such as the Panel Study of Income Dynamics, the National Longitudinal Survey, the Household Income and Labor Dynamics in Australia Study, and the German Socio-Economic Panel.

To shed light on the heterogeneity of the effects by personality traits, I run IV regressions. To make coefficients comparable across regressions, I standardize students' locus of control, growth mindset, and self-control such that their mean is zero and standard deviation is one. The endogenous variables are the change in beliefs from Week 0 to Week 1 and the change in beliefs interacted with the relevant standardized personality trait. The instruments are the treatment dummies and the treatment dummies interacted with the relevant standardized personality trait. The dependent variable is the number of Duolingo lessons completed.

Table 4 presents the results. The odd-numbered columns only control for the baseline beliefs and the even-numbered columns include all the controls. As Columns (1) and (2) show, there is a positive correlation between higher locus of control and higher levels of effort. More importantly, students with a higher locus of control increase their effort more in response to an increase in the perceived productivity of effort. Columns (3) and (4) show that there is neither a detectable correlation between growth mindset and higher levels of effort nor a detectable heterogeneity in the effect of beliefs based on growth mindset. Columns (5) and (6) reveal that there is a positive correlation between higher self control and higher levels of effort. However, there is no detectable heterogeneity in the effect of beliefs based on self control. The heterogeneity in the effect of beliefs by locus of control remains in effect when I allow heterogeneity in the effect of beliefs by growth mindset and self control as shown in Columns (7) and (8).[33]

Why locus of control is an important mediator in the effect of perceived returns to effort on students' effort choices? One potential explanation is that individuals with different levels of locus of control hold different baseline beliefs or update their beliefs differently. I find no evidence of that the baseline beliefs or the change in beliefs are different for students with high levels of locus of control and low levels of locus of control.[34] An alternative explanation is that individuals with different levels of locus of control cope with the same situation differently (Lefcourt, 1992). Upon learning that the returns to Duolingo is lower than expected, students with high levels of locus of control might take some restorative action (i.e. invest effort in a different platform or in a different

[33]These results also hold if I use time spent on Duolingo as the dependent variable. Students with a higher locus of control increase the time spent on Duolingo more in response to an increase in the perceived productivity of effort (statistically significant at the 5% or 10% level). I do not find any statistically significant heterogeneity based on locus of control, growth mindset, or self control in the effect of change in beliefs on dropout.The tables are available upon request.

[34]The tables are available upon request.

Table 4: Heterogeneity of the Effect of Change in Beliefs on Effort by Personality Traits

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Number of Lessons Completed							
Change in Beliefs	2.399	2.566	3.680*	3.899*	4.445*	3.961*	1.229	3.436
	(2.732)	(2.836)	(2.179)	(2.192)	(2.381)	(2.189)	(3.788)	(3.141)
Change in Beliefs*Locus of Control	8.505**	10.95*					8.186*	6.878*
	(4.014)	(6.480)					(4.192)	(3.794)
Locus of Control	9.155**	11.62*					8.731*	8.603**
	(4.191)	(6.100)					(4.740)	(4.138)
Change in Beliefs*Growth Mindset			-1.687	1.561			-0.251	1.799
			(3.034)	(3.396)			(3.878)	(3.709)
Growth Mindset			-0.278	0.579			0.849	1.747
			(1.860)	(2.254)			(2.345)	(2.538)
Change in Beliefs*Self Control					1.148	2.061	-2.028	-0.337
					(2.643)	(2.637)	(3.502)	(2.650)
Self Control					4.847**	4.844*	1.787	2.946
					(2.290)	(2.484)	(3.191)	(3.122)
Constant	29.46***	-113.5	25.73***	-26.77	23.63***	-26.36	31.55***	-53.22
	(6.431)	(179.0)	(4.806)	(105.6)	(5.563)	(116.8)	(8.465)	(143.9)
Controls:								
Baseline Beliefs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Wave Fixed Effect	No	Yes	No	Yes	No	Yes	No	Yes
Demographic	No	Yes	No	Yes	No	Yes	No	Yes
Education	No	Yes	No	Yes	No	Yes	No	Yes
Personality & Preference	No	Yes	No	Yes	No	Yes	No	Yes
Other	No	Yes	No	Yes	No	Yes	No	Yes
Observations	371	371	371	371	371	371	371	371
First Stage F-Stat Eq. 1	6.976	8.163	5.986	7.219	7.121	8.102	4.311	4.806
First Stage F-Stat Eq. 2	2.036	0.835	0.868	0.835	2.783	2.107	1.557	0.891

Notes: 2SLS Regressions. Dependent variable is the number of lessons completed during the four weeks of the experiment. Change in beliefs is the perceived per lesson improvement of the other participants elicited in Week 0 subtracted from the perceived per lesson improvement of the other participants elicited in Week 1. Locus of Control, Growth Mindset and Self Control scores are standardized such that their mean is zero and standard deviation is one. The instruments are the treatment dummies and the treatment dummies interacted with the relevant personality trait(s). Baseline Beliefs are subjects' beliefs about the improvement of other participants for 5 different levels of effort (32, 48, 64, 80, and 96 lessons) elicited in Week 0. See notes of Appendix Table B.1 for detailed description of the control variables. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

activity) to counteract this negative shock and reduce their effort within Duolingo whereas students with low levels of locus of control might feel stuck with Duolingo.

6 Conclusion

How do beliefs about the returns to effort affect effort choices in educational settings? This is an important but under-studied question in economics due to its challenging nature. Students' beliefs about the returns to effort are mostly unknown and endogenous, effort is usually not well measured, and performance measures are a combination of many different types of effort. To overcome these challenges, I conduct a framed field experiment in which perceptions about returns to effort are elicited in an incentivized manner and exogenously manipulated through information interventions, effort is accurately measured, and performance is closely tied to the effort. I then analyze how the changes in students' beliefs causally affect their effort choices.

I find that students' baseline beliefs, on average, are an overestimate of the returns to effort in Duolingo and that students' beliefs move substantially towards the information provided. The latter result documents that students' beliefs are not set in stone and that subtle interventions can manipulate them. I also find that providing anecdotal returns to effort information can be as powerful at changing beliefs as providing information about the average returns to effort. This finding gives a powerful tool to policy makers to nudge beliefs through anecdotal evidence. It also provides suggestive evidence about how individuals establish their beliefs. This finding suggests that a few observations in life may be enough to form beliefs. For example, Altmejd, Barrios-Fernandez, Drlje, Goodman, Hurwitz, Kovac, Mulhern, Neilson and Smith (2020) find that an older sibling's college enrollment and major choice have effects on the younger sibling. One mechanism for their finding is the transmission of information about the college and its returns, which is in line with students learning from anecdotal information.

Using an instrumental variables strategy, I find that students complete more lessons, spend more time, and drop out less in response to an increase in the perceived effectiveness of effort. These findings are in line with substitution effect dominating income effect. These findings are important since they suggest that exogenously manipulating students' beliefs about the effectiveness of effort can be a cost-effective way of eliminating the achievement gaps among students from different socioeconomic backgrounds. I also find that the response of effort to the changes in beliefs is heterogeneous based on one's locus of control. In particular, individuals who believe internal factors determine their outcomes change their effort more in response to the changes in their beliefs.

Finally, we might be worried about the generalizability of the findings of this study to different types of learning environments. The effect of the change in beliefs on effort might be different in compulsory learning settings or in settings where the returns to effort are mostly underestimated. In a recent paper, Rury and Carrell (2020) replicate this paper in a college classroom setting where returns to effort are underestimated and find similar results for the effect of information about the returns to effort on effort, supporting the external validity of the findings of this paper. On the other hand, using a signal-based learning model, Gong (2019) shows that students overestimate the relationship between study time and GPA in a college setting and these high expectations negatively affect study time. Hence, more research in the topic is needed. That being said, in light of the recent experience many schools have had with remote learning due to the spread of COVID-19, results from an online platform are more likely to be externally valid in a future of education that is expected to include blended learning methods.

References

Abramitzky, Ran and Victor Lavy, "How Responsive Is Investment in Schooling to Changes in Redistributive Policies and in Returns?," *Econometrica*, 2014, *82* (4), 1241–1272.

Alan, Sule, Teodora Boneva, and Seda Ertac, "Ever Failed, Try Again, Succeed Better: Results from a Randomized Educational Intervention on Grit," *The Quarterly Journal of Economics*, 2019, *134* (3), 1121–1162.

Almlund, Mathilde, Angela Lee Duckworth, James Heckman, and Tim Kautz, "Personality Psychology and Economics," in Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, eds., *Handbook of the Economics of Education*, Elsevier, 2011, pp. 1 – 181.

Altmejd, Adam, Andres Barrios-Fernandez, Marin Drlje, Joshua Goodman, Michael Hurwitz, Dejan Kovac, Christine Mulhern, Christopher Neilson, and Jonathan Smith, "O Brother, Where Start Thou? Sibling Spillovers on College and Major Choice in Four Countries," *EdWorkingPaper: 20-230*, 2020.

Andreoni, James and Charles Sprenger, "Estimating Time Preferences from Convex Budgets," *American Economic Review*, 2012, *102* (7), 3333–3356.

_ , Michael A. Kuhn, and Charles Sprenger, "Measuring time preferences: A comparison of experimental methods," *Journal of Economic Behavior & Organization*, 2015, *116*, 451–464.

Attanasio, Orazio, Sarah Cattan, Emla Fitzsimons, Costas Meghir, and Marta Rubio-Codina, "Estimating the Production Function for Human Capital: Results from a Randomized Control Trial in Colombia," *American Economic Review*, 2020, *110* (1), 48–85.

Augenblick, Ned, Jesse M. Cunha, Ernesto Dal Bo, and Justin M. Rao, "The Economics of Faith: Using an Apocalyptic Prophecy to Elicit Religious Beliefs in the Field," *Journal of Public Economics*, 2016, *141*, 38–49.

Azmat, Ghazala and Nagore Iriberri, "The importance of relative performance feedback information: Evidence from a natural experiment using high school students," *Journal of Public Economics*, 2010, *94*, 435–452.

_ , Manuel Bagues, Antonio Cabrales, and Nagore Iriberri, "What you don't know... Can't hurt you? A field experiment on relative performance feedback in higher education," *Management Science*, 2019.

Bandiera, Oriana, Valentino Larcinese, and Imran Rasul, "Blissful ignorance? A natural experiment on the effect of feedback on students' performance," *Labour Economics*, 2015, *34*, 13–25.

Barrow, Lisa and Cecilia E. Rouse, "Financial Incentives and Educational Investment: The Impact of Performance-Based Scholarships on Student Time Use," *Education Finance and Policy*, 2018, *13* (4), 419–448.

Bawa, Papia, "Retention in Online Courses: Exploring Issues and Solutions—A Literature Review," *SAGE Open*, 2016, *6* (1).

Becker, Gary, *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education, Second Edition*, 3rd ed., National Bureau of Economic Research, Inc, 1993.

Bénabou, Roland, "The Economics of Motivated Beliefs," *Jean-Jacques Laffont Lecture, Revue d'economie politique 125(5): 665–85.*, 2015.

Bettinger, Eric, "Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores.," *The Review of Economics and Statistics*, 2012, *94* (3), 686–698.

Bishop, John H., "Drinking from the Fountain of Knowledge: Student Incentive to Study and Learn - Externalities, Information Problems and Peer Pressure," in E. A. Hanushek and F. Welch,

eds., *Handbook of Economics of Education*, Vol. 2, Amsterdam: North Holland, 2006, pp. 909–944.

Black, Sandra E., Paul J. Devereux, and Kjell G. Salvanes, "Why the Apple Doesn't Fall Far: Understanding Intergenerational Transmission of Human Capital," *American Economic Review*, March 2005, *95* (1), 437–449.

Bobba, Matteo and Veronica Frisancho, "Learning about oneself: The effects of performance feedback on school choice," *IDB Working Paper No. 728*, 2016.

Boneva, Teodora and Christopher Rauh, "Parental Beliefs about Returns to Educational Investments: The Later the Better?," *Journal of the European Economic Association*, 2018, *16* (6), 1669–1711.

Burks, Stephen V., Jeffrey P. Carpenter, Lorenz Goette, and Aldo Rustichini, "Overconfidence and Social Signalling," *The Review of Economic Studies*, 2013.

Burstyn, Leonardo, Gregory Egorov, and Robert Jensen, "Cool to be Smart or Smart to be Cool? Understanding Peer Pressure in Education," *The Review of Economic Studies*, 2019, *86* (4), 1487–1526.

Clark, Damon, David Gill, Victoria Prowse, and Mark Rush, "Using Goals To Motivate College Students: Theory And Evidence From Field Experiments," *The Review of Economics and Statistics*, 2020, *102* (4), 648–663.

Cobb-Clark, Deborah A., "Locus of control and the labor market," *IZA Journal of Labor Economics*, 2015, *4* (1).

Coleman, James Samuel, Ernest Queener Campbell, and Carol J Hobson, "Equality of educational opportunity," Technical Report, National Center for Educational Statistics 1966.

Conlon, John J., "Major Malfunction: A Field Experiment Correcting Undergraduates' Beliefs about Salaries," *The Journal of Human Resources*, 2019.

Costrell, Robert M., "A Simple Model of Educational Standards," *The American Economic Review*, 1994.

Cunha, Flavio and James Heckman, "The Technology of Skill Formation," *American Economic Review*, May 2007, *97* (2), 31–47.

Cunha, Flávio, Irma Elo, and Jennifer Culhane, "Eliciting Maternal Expectations About the Technology of Cognitive Skill Formation," *NBER Working Paper 19144*, 2013.

Dizon-Ross, Rebecca, "Parents' Beliefs About Their Children's Academic Ability: Implications for Educational Investments," *American Economic Review*, 2019.

Dobrescu, Loretti Isabella, Marco Faravelli, Rigissa Megalokonomou, and Alberto Motta, "Rank Incentives and Social Learning: Evidence from a Randomized Controlled Trial," *IZA Discussion Paper No. 12437*, 2019.

Drerup, Tilman, Benjamin Enke, and Hans-Martin von Gaudecker, "Measurement Error in Subjective Expectations and the Empirical Content of Economic Models," *IZA Discussion Paper No. 8535*, 2014.

Duckworth, Angela L., *Grit: The Power of Passion and Perseverance*, New York: Scribner, 2016.

Dweck, Carol S., *Mindset: The New Psychology of Success*, Random House, 2006.

Eil, David and Justin M. Rao, "The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself," *American Economic Journal: Microeconomics*, May 2011, *3* (2), 114–38.

Ersoy, Fulya, "Returns to Effort: Experimental Evidence from an Online Language Platform," *Experimental Economics*, 2020.

Figlio, David, Paolo Giuliano, Umut Ozek, and Paola Spaienza, "Long-term Orientation and Educational Performance," *American Economic Journal: Economic Policy*, 2019, *11* (4).

Findley, Maureen J. and Harris M. Cooper, "Locus of Control and Academic Achievement: A Literature Review," *Journal of Personality and Social Psychology*, 1983, *44* (2), 419–427.

Fraja, Gianni De, Tania Oliveira, and Luisa Zanchi, "Must Try Harder: Evaluating The Role Of Effort In Educational Attainment," *The Review of Economics and Statistics*, 2010, *92* (3), 577–597.

Fryer, Roland G., "Financial Incentives and Student Achievement: Evidence From Randomized Trials," *The Quarterly Journal of Economics*, 2011.

_ , "The Production of Human Capital in Developed Countries: Evidence from 196 Randomized Field Experiments," in "Handbook of Field Experiments," Vol. 2, Amsterdam: North Holland, 2017, pp. 95–322.

Gächter, Simon and Elke Renner, "The effects of (incentivized) belief elicitation in public goods experiments," *Experimental Economics*, 2010, *13* (3), 364–377.

Gneezy, Uri, John List, Jeffrey A. Livingston, Sally Sadoff, Xiangdong Qin, and Yang Xu, "Measuring Student Success: The Role of Effort on the Test Itself," *American Economic Review: Insights*, 2019, *1* (3), 291–308.

Gong, Yifan, "Signal-based Learning Models without the Rational Expectations Assumption: Identification and Counterfactuals," *Working Paper*, 2019.

Harrison, Glenn W. and John A. List, "Field Experiments," *Journal of Economic Literature*, 2004, *42*, 1009–1055.

Hastings, J. S., C. A. Neilson, and S. D. Zimmerman, "The Effects of Earnings Disclosure on College Enrollment Decisions.," *NBER Working Paper 21300.*, 2015.

Heckman, James and Tim Kautz, "Fostering and Measuring Skills: Interventions that Improve Character and Cognition," in James Heckman, J.E. Humphries, and Tim Kautz, eds., *In The Myth of Achievement Tests: The GED and the Role of Character in American Life*, University of Chicago Press, 2014, pp. 341–430.

Hirshleifer, Sarojini R., "Incentives for Effort or Outputs? A Field Experiment to Improve Student Performance," *Working Paper*, 2016.

Jensen, Robert, "The (Perceived) Returns to Education and the Demand for Schooling," *The Quarterly Journal of Economics*, 2010, pp. 515–548.

Lefcourt, H. M., "Durability and impact of the locus of control construct.," *Psychological Bulletin*, 1992, *112* (3), 411–414.

Levitt, Steven D., John A. List, Susanne Neckermann, and Sally Sadoff, "The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance," *American Economic Journal: Economic Policy*, November 2016, *8* (4), 183–219.

Mobius, Markus, Muriel Niederle, Paul Niehaus, and Tanya Rosenblat, "Managing Self-Confidence: Theory And Experimental Evidence," *NBER Working Paper 17014*, 2011.

Nguyen, Trang, "Information, Role Models and Perceived Returns to Education: Experimental Evidence from Madagascar," *Working Paper*, 2008.

Owen, John D., *Why Our Kids Don't Study? An Economist's Perspective*, The John Hopkins University Press, 1995.

Paunesku, David, Gregory M. Walton, Carissa Romero, Eric N. Smith, David S Yeager, and Carol S. Dweck, "Mind-Set Interventions are a scalable treatment for academic underachievement," *Psychological Science*, 2015, *26*, 784–793.

Rammstedt, Beatrice and Oliver P. John, "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*, 2007, *41* (1), 203 – 212.

Rosen, Sherwin, "Human Capital: A Survey of Empirical Research," *Research in Labor Economics*, 1977, *1*, 3–40.

Rotter, J. B., "Generalized Expectancies for Internal versus External Control of Reinforcement.," *Psychological Monographs: General and Applied*, 1966, *80*, 1–28.

Rotter, Julian B., "Internal versus external control of reinforcement: A case history of a variable.," *American Psychologist*, 1990, *45* (4), 489–493.

Rury, Derek and Scott Carrell, "Knowing What it Takes: The Effect of Information About Returns To Studying on Study Effort and Achievement," *Working Paper*, 2020.

Schotter, Andrew and Isabel Trevino, "Belief Elicitation in the Laboratory," *Annual Review of Economics*, 2014, *6* (103-128).

Schultz, P. W., J. M. Nolan, R. B. Cialdini, N. J. Goldstein, and V. Griskevicius, "The constructive, destructive, and reconstructive power of social norms.," *Psychological Science*, 2007, *18*, 429–434.

Sharot, Tali and Neil Garrett, "Forming Beliefs: Why Valence Matters," *Trends in Cognitive Sciences*, 2016, *20* (1), 25–33.

Stinebrickner, Ralph and Todd R. Stinebrickner, "The causal effect of studying on academic performance," *The B.E. Journal of Economic Analysis and Policy*, 2008.

_ **and** _ , "Academic Performance and College Dropout: Using Longitudinal Expectations Data to Estimate a Learning Model," *Journal of Labor Economics*, 2014, *32* (3), 601–644.

Stinebrickner, Todd and Ralph Stinebrickner, "Learning about Academic Ability and the College Dropout Decision," *Journal of Labor Economics*, 2012, *30* (4), 707–748.

Tangney, June P., Roy F. Baumeister, and Angie Luzio Boone, "High Self-Control Predicts Good Adjustment, Less Pathology, Better Grades, and Interpersonal Success," *Journal of Personality*, 2004, *72* (2), 271–324.

Thaler, Richard H., "The Winner's Curse," *The Journal of Economic Perspectives*, 1988, *2* (1), 191–202.

Wang, Stephanie W., "Incentive effects: The case of belief elicitation from individuals in groups," *Economics Letters*, 2011, *111* (1), 30 – 33.

Wiswall, Matthew and Basit Zafar, "Determinants of College Major Choice: Identification using an Information Experiment," *The Review of Economic Studies*, 2015, *82* (2), 791–824.

APPENDIX

Effects of Perceived Productivity on Study Effort: Evidence from a Field Experiment

Fulya Ersoy

A Experimental Design Details and Materials

A.1 Subject Recruitment

For this study, I recruited students who are intrinsically motivated to learn a language. I first decided the set of universities from which I recruit the subjects. Using College Navigator from the National Center for Education Statistics, I restricted my attention to schools that are public, offer at least a Bachelor's degree, have at least 20,000 undergraduates and admit 30% to 70% of its applicants. This search gave me 54 universities. Then, I further restricted my attention to the universities in California which gave me a list of 15 universities. One of these universities did not allow me to recruit its students, which left me with the following list of 14 Californian universities: 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, San Francisco State University, and San Jose State University.

To recruit students from these universities, I collected the email addresses of faculty and staff members in each department (17,166 email addresses in total). I sent an email to these individuals asking their help disseminating my recruitment flyer (see Appendix Figure A.7 for a sample flyer) to the students in their departments.[35] The flyer specifically stated that the students who want to participate in the language learning research study should be motivated to learn Spanish and should know no or little Spanish to start with. After students signed up, they took an eligibility

[35]I thank the faculty and staff members who help me share the flyers with their students.

check survey. If they failed one of the eligibility criteria, they could not become a participant in this study.[36]

A.2 Spanish Assessment Test Details

To construct the test, I scrape all practice tasks from all Duolingo lessons from the Beginner and Intermediate Skills. I create two separate tests based on these lessons, Test A and Test B.[37] The order of the tests is randomized across participants. That is, participants are randomly assigned to receive Test A or Test B as the initial test. Those who took Test A as the initial test received Test B as the final test, and vice versa. Each test has different types of questions. The majority of the questions are multiple choice questions that ask participants to select only one answer choice from a list of choices (question type S) or to select one or more answer choices from a list of choices (question type M). There are also questions where participants need to determine whether a noun is masculine or feminine (question type MF) and mix & match type of questions where the subject needs to match the Spanish words with their English counterparts or match the conjugations of a verb with the pronouns (question type MM). Each test is checked for correctness by at least 3 native Spanish speakers.[38]

Scoring of the test is done such that each Duolingo lesson has an equal weight on test score, subject to a rounding error. Since, Intermediate Skills set has almost twice as many lessons as Beginner Skills set and an equal number of questions from both skill sets are in the Spanish test, questions based on the lessons from Intermediate Skills set are worth almost twice as much as questions based on the lessons from Beginner Skills set. A subject receives $1 - Pr(\textit{randomly guessing correctly})$ points if he answers a question that belongs to a beginner lesson correctly and $-Pr(\textit{randomly guessing correctly})$ points if he answers a question that belongs to a beginner lesson incorrectly. Similarly, a subject receives $2(1 - Pr(\textit{randomly guessing correctly}))$ if he answers a question that belongs to an intermediate lesson correctly and $-2Pr(\textit{randomly guessing correctly})$ points if he answers a question that belongs to an intermediate lesson incorrectly. $Pr(\textit{randomly guessing correctly})$ depends on

[36] The eligibility criteria are being a student aged 18 or above, being highly interested in learning Spanish, being able to commit up to 4 hours of studying Spanish online, and knowing no or little Spanish. These criteria helped me recruit motivated individuals and increase the homogeneity of the sample to increase the statistical power.

[37] 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).

[38] 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.

the type of question and the number of answer choices available.[39] With this scoring rule, randomly guessing in all questions would result in a score of 0 on average and correctly answering all questions would result in a maximum score of 120. To have a finer grid, I renormalize the test scores such that the maximum score is 1000.

[39]For example, it is $\frac{1}{NumberOfAnswerChoices}$ for question type S, it is $\frac{1}{(2^{NumberOfAnswerChoices})-1}$ for question type M, etc.

A.3 Information about Duolingo Lessons

There are in total 317 lessons in a Duolingo tree. 67 of these lessons belong to Beginner Skills, 142 of these lessons belong to Intermediate Skills and 108 of these lessons belong to Advanced Skills. Each Duolingo lesson takes 5 to 10 minutes to complete.

To see what these skill levels are, you can have a look at the list below.

Beginner		**Intermediate**		**Advanced**	
Skills	**Lessons**	**Skills**	**Lessons**	**Skills**	**Lessons**
Basics 1	1-3	Verbs: Present 2	1-7	Verbs: Past Perfect	1-5
Common Phrases	1-3	Determiners	1-3	Nature	1-7
Basics 2	1-4	Adverbs	1-10	Vocabulary 2	1-8
Food	1-4	Objects	1-8	Verbs: Infinitive 2	1-10
Animals	1-4	To Be: Ser/Estar	1-2	Medical	1-9
Possessives	1-3	Places	1-9	Verbs: Gerund	1-4
Clothing	1-4	People	1-4	Business	1-5
Questions	1-2	Object Pronouns	1-5	Communication	1-3
Verbs: Present 1	1-3	Numbers	1-5	Verbs: Haber	1
Food 2	1-4	Verbs: Past	1-10	Spiritual	1-2
Family	1-4	Verbs: Present 3	1-10	Verbs: Future	1-6
Sizes	1	Verbs: Infinitive 1	1-10	Arts	1-5
Household	1-6	Verbs: Phrasal Future Tense	1-10	Verbs: Future Perfect	1-2
Occupations	1-6	Countries	1-4	Politics	1-10
Dates and Time	1-7	Adjectives 2	1-10	Verbs: Subjunctive/Imperative	1-3
Adjectives 1	1-9	Pronouns	1-3	Science	1-5
		Directions	1-3	Verbs: Modal	1-2
		Education	1-4	Verbs: Conditional	1
		Vocabulary 1	1-10	Verbs: Past Imperfect	1-5
		Verbs: Participle	1-2	Verbs: Subjunctive Past	1
		Feelings	1-3	Vocabulary 3	1-10
		Verbs: Present Perfect	1-10	Verbs: Conditional Perfect	1-2

A.4 Information about the Test

We want to give you some information about the test.

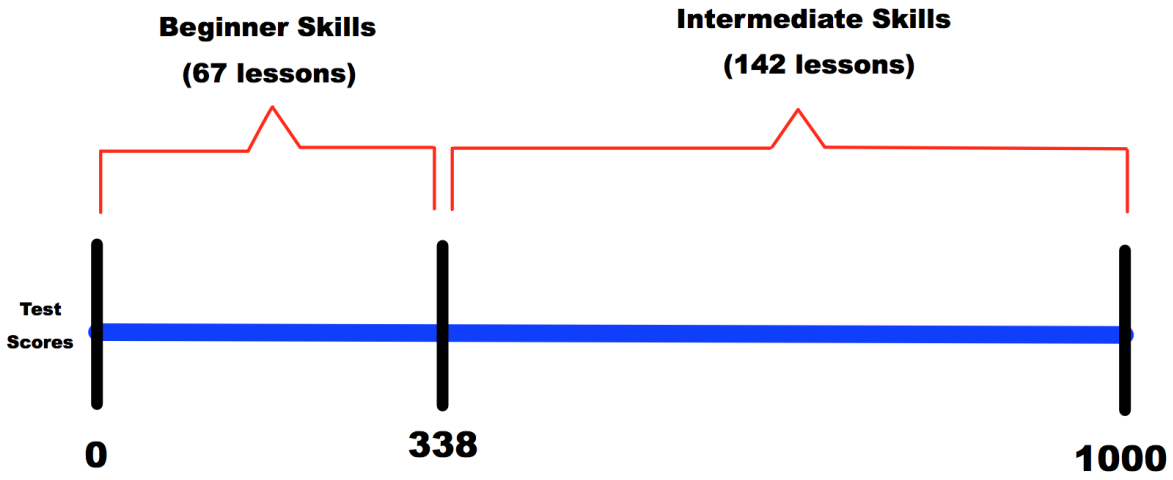
This test consists of 80 questions in total as you know. 40 of these questions are based on Duolingo Beginner skills and 40 of these questions are based on Duolingo Intermediate skills.

A person who has perfectly mastered Duolingo Beginner skills (the first 67 lessons) and nothing else should be able to correctly answer the first 40 questions of the test and get a score of 338 points.

A person who has perfectly mastered Duolingo Beginner skills (the first 67 lessons) and Duolingo Intermediate skills (the next 142 lessons) and nothing else should be able to correctly answer all 80 questions of the test and get a score of 1000 points.

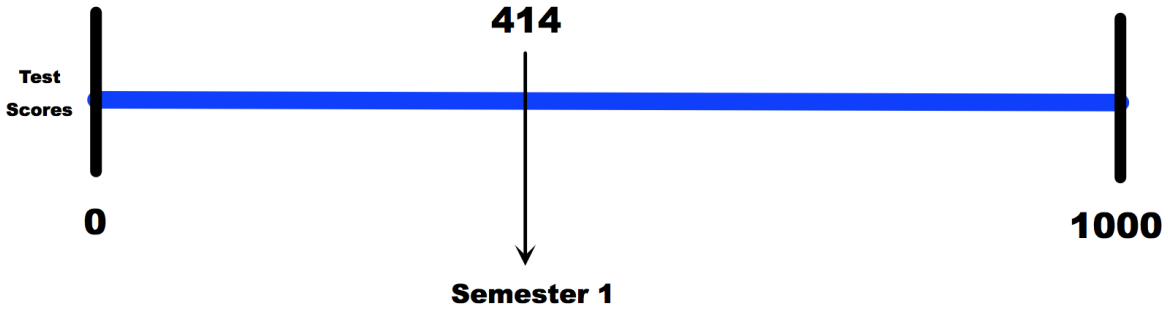
You can see the graphical representation below.

Scoring rule was determined such that each Duolingo lesson has an equal weight on test score. Since there are twice as many lessons in intermediate skills as there are lessons in beginner skills, questions based on the lessons from intermediate skills are worth twice as much as questions based on the lessons from beginner skills.



To be able to give you more information about how relevant this test for measuring Spanish knowledge, we gave the test to the students who are currently taking college level Spanish courses and collected data on their test scores.

We find that students who finished Semester 1 in a college level Spanish course would get a score of 414 on average from the test. You can see the graphical representation below.

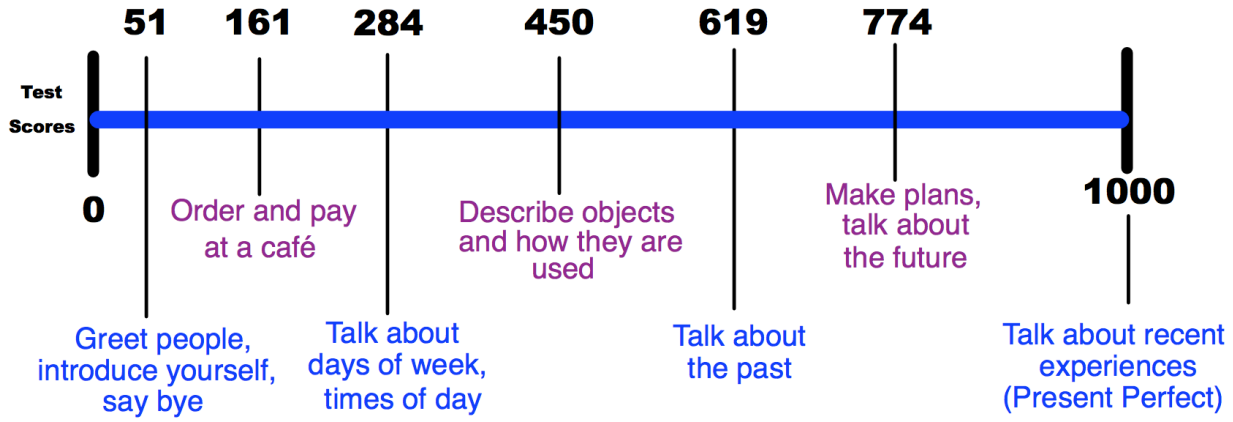


A semester approximately corresponds to 5-hour per week college level Spanish courses for 12 weeks. To grasp what finishing the first semester mean, let's look at some examples of what the students who finish Semester 1 can do.

Semester 1: Students who complete Semester 1 can

- understand the main point of short simple texts (such as messages on postcards, directions and descriptions)
- ask for or pass on personal details in written form
- ask and answer simple questions about themselves, where they live, people they know and things they have
- introduce themselves and express gratitude
- identify familiar words and phrases in a speech

Finally, we believe the figure below will help you get a better understanding of the scores of the test.

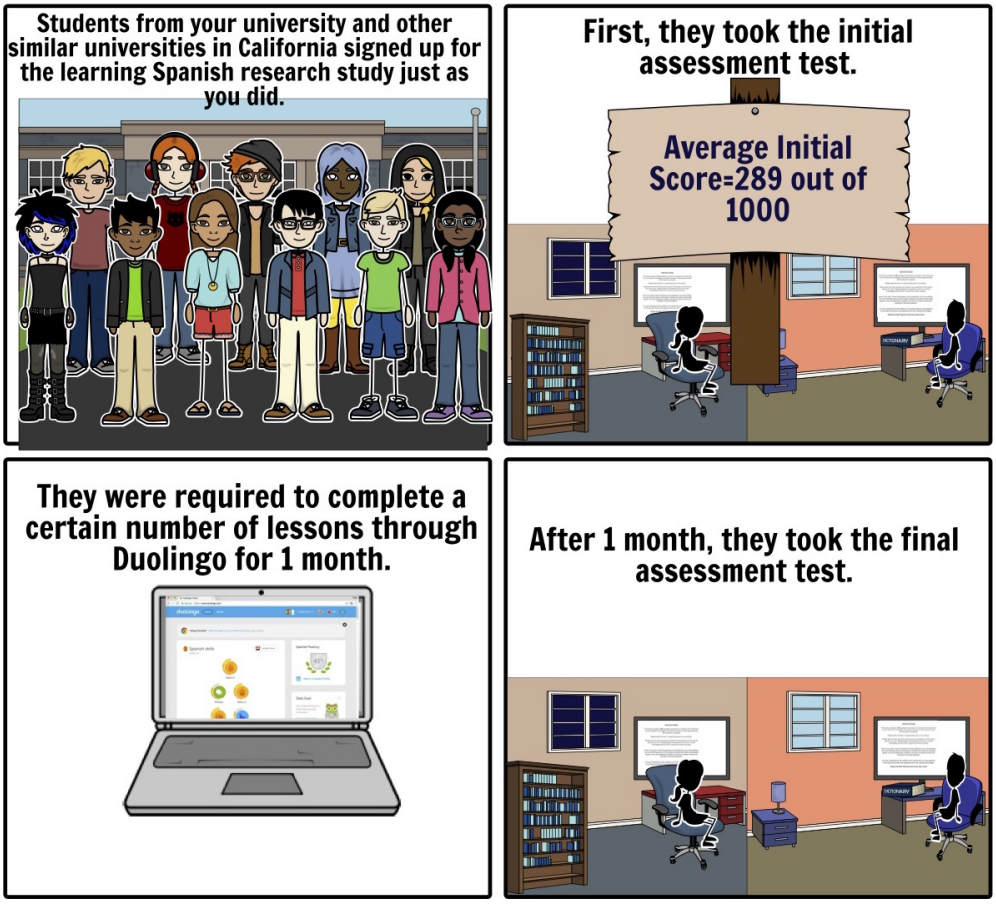


A.5 Information about Other Participants

The participants of the previous study first took the initial test. The average score at the initial test was 289 points out of 1000 points. Then, they were randomly assigned to one of the following assignment groups:

- 8 Duolingo lessons per week,
- 12 Duolingo lessons per week,
- 16 Duolingo lessons per week,
- 20 Duolingo lessons per week, and
- 24 Duolingo lessons per week.

The participants in the 8 lessons per week group were assigned to complete 8 Duolingo lessons per week, the participants in the 12 lessons per week group were assigned to complete 12 Duolingo lessons per week, etc. Each Duolingo lesson takes 5 to 10 minutes to complete.
More than 90% of the participants stated that they only studied Spanish with Duolingo during the study period.



A.6 Benchmarking Exercise

To be able to provide the participants with information on the meaning of the test scores, I do a benchmarking exercise. I contacted all the Spanish departments of the universities in my sample to recruit students from the Beginner, Elementary or Intermediate classes to take my Spanish test. Only UCSD and UCI allowed me to reach out their students who are enrolled in a Spanish course. I also recruited some students from Stanford through hardcopy flyers. I then sent these students the test along with a small survey that asks which Spanish course they are currently taking (Beginner, Elementary or Intermediate) and how many hours of Spanish they have taken for the current term. In total, 36 students took my survey and the test (75% of them are from UCSD).

Using test scores and the answers to the two survey questions, I benchmark the test scores. From the survey questions, I construct the hours attended variable. For example, for a student who is taking the Elementary Spanish course at UCI and who has attended 10 hours of instruction for the current term, hours attended is equal to $5\ hours/week * 10\ weeks + 10\ hours = 60\ hours$.

Similarly, for a student who is taking the Intermediate Spanish at UCSD and who has attended 10 hours of instruction for the current term, hours attended is equal to 5 $hours/week * 12$ $weeks *$ 2 $semesters + 10$ $hours = 130$ $hours$. Then, I regress test scores on hours attended. I find that the constant is 273.94 and the slope coefficient is 2.34, both statistically significant at the 1% level. Hence, I calculate the average test score for a student who has just finished the first semester of a college level Spanish course which has 60 hours of instruction (5 hours/week 12 weeks) as 414 out of 1000 which is what I report to the participants in my experiment.

A.7 Eliciting Time and Risk Preferences

To elicit preferences, I use the Convex Time Budget (CTB) method and follow the procedures in Andreoni et al. (2015). This method uses variation in linear budget constraints over early and later payments to identify the convexity of preferences. In particular, participants make 24 decisions over how to allocate money between two points in time. Each decision features a series of options. Each option consists of a sooner payment and a later payment with varying amounts. Table A.1 summarizes the parameters of each decision. For example, in the case of Decision 1, the sooner payment is today, the later payment is in 2 weeks and the price ratio is 1. That gives rises the following 6 options one of which the subject needs to choose: ($5 today, $0 in 2 weeks), ($4 today, $1 in 2 weeks), ($3 today, $2 in 2 weeks), ($2 today, $3 in 2 weeks), ($1 today, $4 in 2 weeks) and ($0 today, $5 in 2 weeks). Subjects are asked to pick their favorite option in each decision. At the end, one of the decisions is randomly chosen for payment. The payment subjects get varies between $2.5 and $5 depending on their choices and luck. Subjects get paid through Google Wallet.[40]

Table A.1: Parameters used in the Convex Time Budget Method

t(days until first payment)	k(delay)	P(price ratio): $Px_t + x_{t+k} = 5$
0 days	14 days	$1,\ \frac{10}{9},\ \frac{10}{8},\ \frac{10}{7},\ \frac{10}{6},\ 2$
0 days	28 days	$1,\ \frac{10}{9},\ \frac{10}{8},\ \frac{10}{7},\ \frac{10}{6},\ 2$
14 days	14 days	$1,\ \frac{10}{9},\ \frac{10}{8},\ \frac{10}{7},\ \frac{10}{6},\ 2$
14 days	28 days	$1,\ \frac{10}{9},\ \frac{10}{8},\ \frac{10}{7},\ \frac{10}{6},\ 2$

x_t is the sooner payment and x_{t+k} is the later payment.

[40] Completion payment and bonus payments are paid as Amazon e-gift cards. I choose to make the Preference Survey payments through Google Wallet instead of Amazon e-gift cards in order to equate transaction costs of sooner and later payments.

By collecting the choices of participants across 24 decision problems and assuming a utility function in the form below, one can identify the key parameters of the utility function: curvature (α), discounting (δ) and time inconsistency parameter (β).

$$U(x_t, x_{t+k}) = \begin{cases} x_t^\alpha + \beta\delta^k x_{t+k}^\alpha & \text{if } t = 0 \\ x_t^\alpha + \delta^k x_{t+k}^\alpha & \text{if } t > 0 \end{cases}$$

By re-arranging the intertemporal Euler equation for the solution of this problem and assuming an additive error structure, I estimate Equation 6 at the individual level and recover the key parameters of the utility function by non-linear transformations of the regression coefficients.[41] By using the variation in the interest rate, I identify the curvature of the utility function. For a fixed interest rate, I identify discounting parameter by using the variation in delay (whether it is 2 weeks or 4 weeks). Finally, I use the variation in the date of sooner payment to identify time inconsistency parameter.

$$ln(\frac{x_t}{x_{t+k}}) = \frac{ln(\beta)}{\alpha - 1}\mathbb{1}(t = 0) + \frac{ln(\delta)}{\alpha - 1}k + \frac{1}{\alpha - 1}ln(P) \tag{6}$$

[41] If x_t or x_{t+k} is equal to \$0, I change it to \$0.01 so that the log ratio is well-defined.

A.8 Experimental Figures

Figure A.1: Example Practice Tasks from a Duolingo Lesson

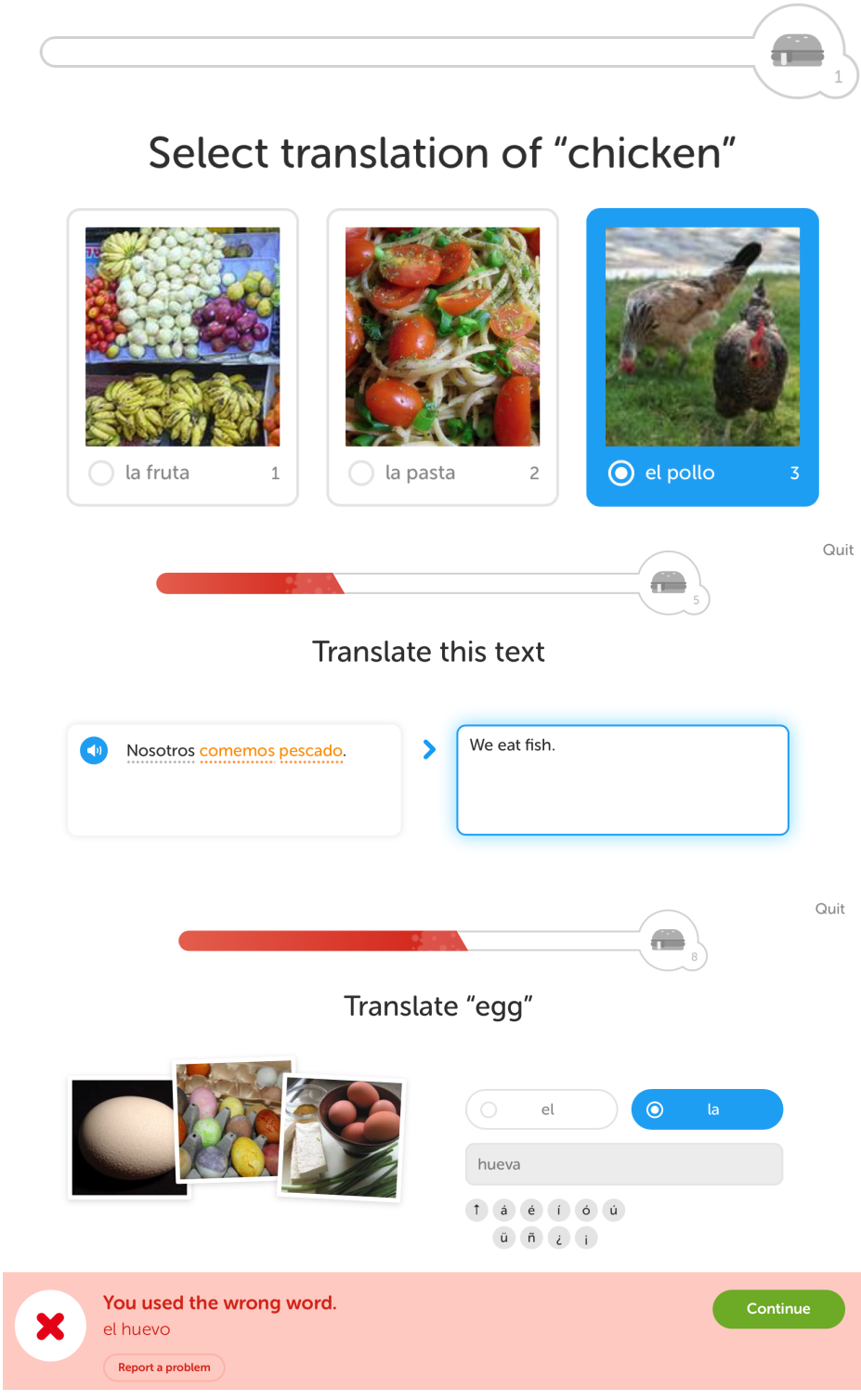


Figure A.2: List of Duolingo Lessons

Beginner		Intermediate		Advanced	
Skills	**Lessons**	**Skills**	**Lessons**	**Skills**	**Lessons**
Basics 1	1-3	Verbs: Present 2	1-7	Verbs: Past Perfect	1-5
Common Phrases	1-3	Determiners	1-3	Nature	1-7
Basics 2	1-4	Adverbs	1-10	Vocabulary 2	1-8
Food	1-4	Objects	1-8	Verbs: Infinitive 2	1-10
Animals	1-4	To Be: Ser/Estar	1-2	Medical	1-9
Possessives	1-3	Places	1-9	Verbs: Gerund	1-4
Clothing	1-4	People	1-4	Business	1-5
Questions	1-2	Object Pronouns	1-5	Communication	1-3
Verbs: Present 1	1-3	Numbers	1-5	Verbs: Haber	1
Food 2	1-4	Verbs: Past	1-10	Spiritual	1-2
Family	1-4	Verbs: Present 3	1-10	Verbs: Future	1-6
Sizes	1	Verbs: Infinitive 1	1-10	Arts	1-5
Household	1-6	Verbs: Phrasal Future Tense	1-10	Verbs: Future Perfect	1-2
Occupations	1-6	Countries	1-4	Politics	1-10
Dates and Time	1-7	Adjectives 2	1-10	Verbs: Subjunctive/Imperative	1-3
Adjectives 1	1-9	Pronouns	1-3	Science	1-5
		Directions	1-3	Verbs: Modal	1-2
		Education	1-4	Verbs: Conditional	1
		Vocabulary 1	1-10	Verbs: Past Imperfect	1-5
		Verbs: Participle	1-2	Verbs: Subjunctive Past	1
		Feelings	1-3	Vocabulary 3	1-10
		Verbs: Present Perfect	1-10	Verbs: Conditional Perfect	1-2

Figure A.3: Information provided across Treatments

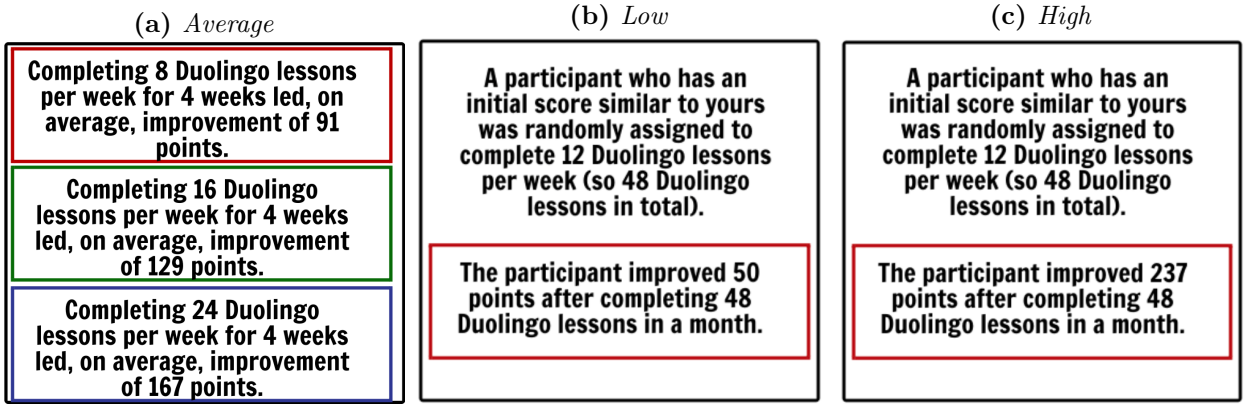


Figure A.4: An Example Information Intervention:*Average Information* Treatment

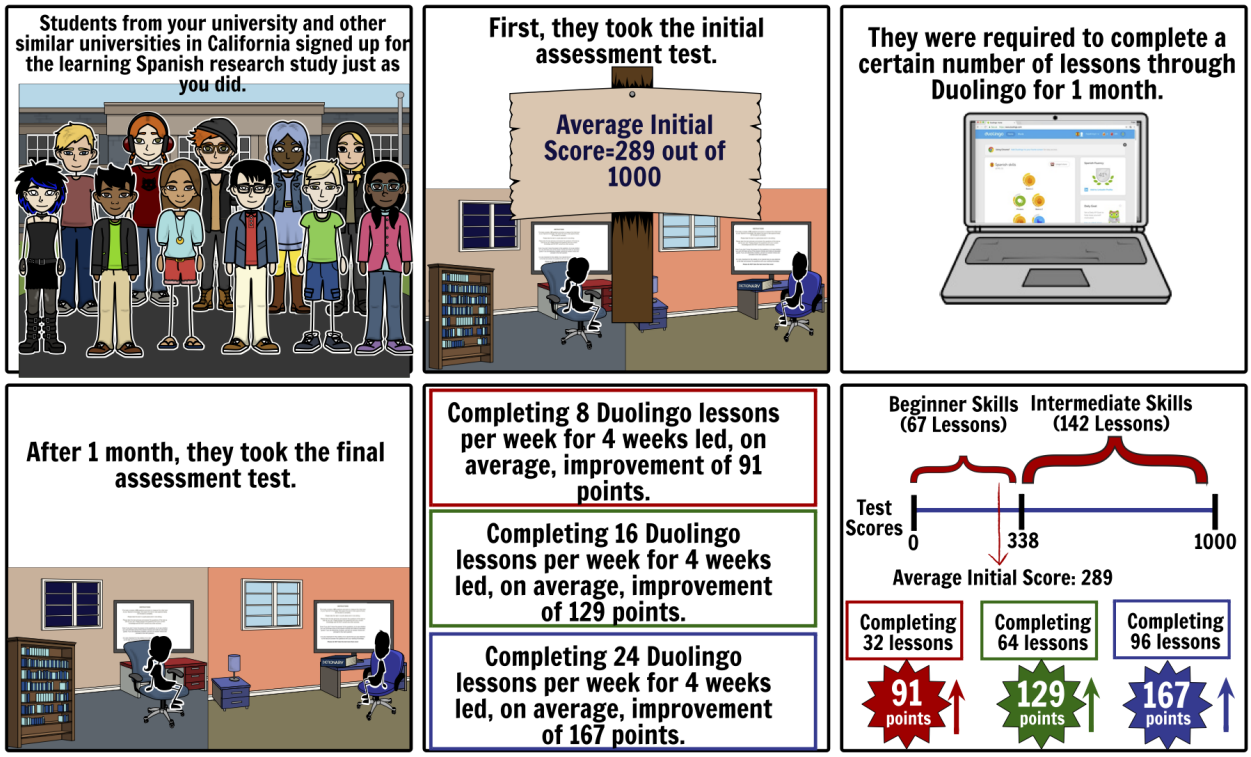
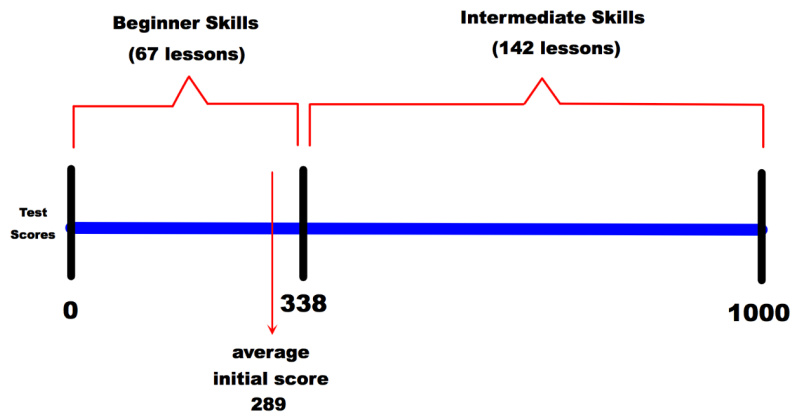


Figure A.5: Belief Questions about the Other Participants

Now, we will ask your guess about the improvement of the participants of the previous round.

The better your guess for the following questions, the higher your bonus payment will be. Please answer the questions as well as you can.



On average, how much improvement in test sores do you think results from completing **8** lessons per week for 4 weeks (**32** lessons in total)?

By improvement, we mean the difference between the final test score and the initial test score. (Please only type in numbers, no words or blanks!)

On average, how much improvement in test sores do you think results from completing **12** lessons per week for 4 weeks (**48** lessons in total)?

By improvement, we mean the difference between the final test score and the initial test score. (Please only type in numbers, no words or blanks!)

On average, how much improvement in test sores do you think results from completing **16** lessons per week for 4 weeks (**64** lessons in total)?

By improvement, we mean the difference between the final test score and the initial test score. (Please only type in numbers, no words or blanks!)

On average, how much improvement in test sores do you think results from completing **20** lessons per week for 4 weeks (**80** lessons in total)?

By improvement, we mean the difference between the final test score and the initial test score. (Please only type in numbers, no words or blanks!)

On average, how much improvement in test sores do you think results from completing **24** lessons per week for 4 weeks (**96** lessons in total)?

By improvement, we mean the difference between the final test score and the initial test score. (Please only type in numbers, no words or blanks!)

(If one of these questions is randomly chosen for payment, we will discount $0.50 from your bonus payment for each point by which your guess and the correct answer to the question chosen differs.)

To review the information you have already seen, you can click on the links below.

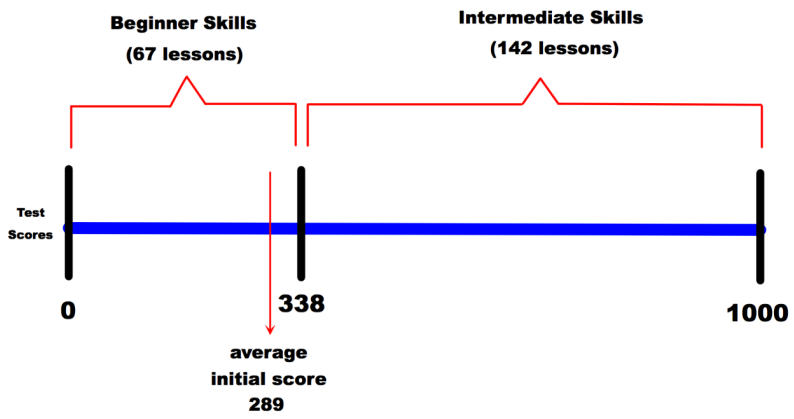
What Scores Mean | Overview of Duolingo Lessons | Bonus Payment

Figure A.6: Belief Questions about the Self

Now, we will ask your guess about what your improvement would be if you were a participant in this previous round.

Your answer to the following questions will NOT affect your payments from the study. Please answer the questions as well as you can.

Imagine that you participated in the previous round of the study and your score at the initial test was **289** points out of **1000** points.



What do you think your improvement in test scores would be if you complete **8** lessons per week for 4 weeks (**32** lessons in total)?

By improvement, we mean the difference between the final test score and the initial test score. (Please only type in numbers, no words or blanks!)

What do you think your improvement in test scores would be if you complete **12** lessons per week for 4 weeks (**48** lessons in total)?

By improvement, we mean the difference between the final test score and the initial test score. (Please only type in numbers, no words or blanks!)

What do you think your improvement in test scores would be if you complete **16** lessons per week for 4 weeks (**64** lessons in total)?

By improvement, we mean the difference between the final test score and the initial test score. (Please only type in numbers, no words or blanks!)

What do you think your improvement in test scores would be if you complete **20** lessons per week for 4 weeks (**80** lessons in total)?

By improvement, we mean the difference between the final test score and the initial test score. (Please only type in numbers, no words or blanks!)

What do you think your improvement in test scores would be if you complete **24** lessons per week for 4 weeks (**96** lessons in total)?

By improvement, we mean the difference between the final test score and the initial test score. (Please only type in numbers, no words or blanks!)

To review the information you have already seen, you can click on the links below.

What Scores Mean | Overview of Duolingo Lessons

Figure A.7: Recruitment Flyer

ARE YOU INTERESTED IN LEARNING *SPANISH* ONLINE THIS SUMMER?

PARTICIPATE IN THIS RESEARCH STUDY!

HELP US LEARN MORE ABOUT LANGUAGE LEARNING!

LEARN SPANISH ONLINE!

GET COMPENSATION!

AM I THE RIGHT PERSON FOR THIS STUDY?

YOU ARE IF

- You don't know any Spanish or you know very little Spanish
- You are motivated to learn (more) Spanish
- You can commit studying Spanish online for **up to 4 hours per week** for **1 month** starting from mid June
- You are at least 18 years old
- You are a student at SDSU

COMPENSATION

Upon your successful completion of the study, you will get **$50** (on average) which will be paid to you as an Amazon Gift Card.

CONTACT INFORMATION

If interested in participating, please provide your e-mail address through this survey form:

http://tinyurl.com/startStudyingSpanish

If you have any questions, please contact us
start.studying.spanish@gmail.com

This research is approved by Stanford Panel on Non-Medical Human Subjects. The IRB protocol number is 36512.

B Tables

Table B.1: Summary Statistics for the Baseline Sample

	Averages								P-Values		
					Treatments						
	(1) Initial Sample	(2) Wave 1 Sample	(3) Wave 2 Sample	(4) Baseline Sample	(5) *No Information*	(6) *Average Information*	(7) *Low*	(8) *High*	(9) Wave 1 vs Wave 2	(10) Initial vs Baseline	(11) Joint Test across Treatments
Demographic Controls:											
Female	.69	.72	.67	.71	.72	.66	.73	.71	.174	.418	.636
African-American	.05	.07	.04	.05	.06	.02	.07	.05	.2	.671	.395
Asian	.4	.31	.49	.44	.35	.52	.46	.44	0	.02	.119
Caucasian	.34	.38	.31	.34	.44	.3	.28	.35	.048	.81	.081
Hispanic	.1	.13	.07	.08	.09	.07	.11	.06	.007	.122	.598
Age	23.08	23.74	22.54	23.01	23.76	23.36	22.45	22.46	.002	.635	.15
Education Controls:											
Arts, Humanities, and Soc. Sci.	.27	.33	.22	.25	.22	.24	.28	.26	.003	.118	.86
Business and Economics	.06	.1	.02	.04	.03	.04	.04	.06	0	.035	.74
Health and STEM	.59	.49	.68	.63	.65	.62	.61	.66	0	.006	.876
Freshman	.1	.07	.13	.1	.06	.05	.16	.14	.018	.967	.02
Sophomore	.14	.12	.17	.17	.14	.2	.12	.23	.078	.005	.169
Junior	.25	.24	.25	.23	.26	.27	.22	.18	.905	.351	.416
Senior	.3	.37	.24	.27	.27	.23	.28	.31	.001	.061	.659
GPA	3.39	3.34	3.43	3.42	3.45	3.41	3.42	3.38	.013	.056	.731
SAT/ACT Math	.08	-.02	.17	.13	.14	.14	.1	.14	0	.012	.952
SAT/ACT Reading	.07	.01	.13	.09	.13	.08	.09	.07	.012	.3	.9
SAT/ACT Missing	.5	.57	.43	.48	.5	.47	.46	.49	0	.304	.947
Mother's Education	14.73	14.52	14.9	14.58	14.61	14	15	14.73	.231	.237	.356
Father's Education	14.4	14.2	14.57	14.69	14.69	14.34	15.12	14.6	.329	.048	.704
Initial Test Score	-	-	-	306.14	326.31	303.54	308.27	286.06	-	-	.322
Personality Controls:											
Locus of Control	3.6	3.74	3.48	3.58	3.66	3.75	3.39	3.52	.042	.691	.377
Growth Mindset	28.9	28.75	29.03	28.86	29.66	28.33	28.35	29.1	.343	.696	.021
Self Control	44.05	43.82	44.23	44.51	44.8	44.72	44.46	44.05	.524	.065	.923
Extroversion	6.25	6.27	6.23	6.22	6.16	6.25	6.31	6.15	.779	.622	.922
Agreeableness	7.24	7.4	7.11	7.27	7.37	7.26	7.31	7.15	.018	.542	.764
Neuroticism	5.99	6.02	5.97	5.99	5.92	5.87	6.16	6.02	.756	.996	.702
Conscientiousness	7.41	7.36	7.45	7.39	7.36	7.34	7.36	7.49	.444	.619	.898
Openness	7.43	7.51	7.35	7.37	7.39	7.45	6.96	7.67	.228	.235	.021

Table B.1: Summary Statistics for the Baseline Sample

	Averages								P-Values		
					Treatments						
	(1) Initial Sample	(2) Wave 1 Sample	(3) Wave 2 Sample	(4) Baseline Sample	(5) *No Information*	(6) *Average Information*	(7) *Low*	(8) *High*	(9) Wave 1 vs Wave 2	(10) Initial vs Baseline	(11) Joint Test across Treatments
Preference Survey Controls:											
α (Risk Aversion)	-	-	-	.57	.68	.11	.78	.7	-	-	.14
β (Time Inconsistency)-	-	-	-	1.04	1.05	1.04	1.03	1.04	-	-	.922
δ (Patience)	-	-	-	.9991	1.001	.9980	.9983	.9991	-	-	.171
Preference Survey Missing	-	-	-	.09	.1	.11	.08	.08	.-	-	.872
Other Controls:											
Commitment	.83	.79	.86	.84	.85	.87	.82	.84	.018	.273	.795
Interest	.89	.88	.9	.88	.86	.88	.89	.89	.274	.134	.913
Don't Know Spanish	.44	.39	.49	.46	.4	.43	.45	.56	.009	.298	.118
Know Beginner Spanish	.49	.54	.45	.48	.54	.48	.49	.4	.028	.434	.24
Know Elementary Spanish	.06	.07	.05	.06	.05	.08	.05	.04	.441	.718	.66
Native English Speaker	.66	.73	.6	.65	.76	.59	.67	.6	.001	.679	.053
Serious Learner	.52	.52	.53	.52	.52	.58	.46	.54	.831	.897	.42
Income	96743	90088	102238	97736	91173	97474	94719	107786	.16	.767	.749
Expenditure	293.55	280.81	304.07	284.14	261.22	255.3	305.36	315.63	.346	.325	.333
Observations	628	284	344	391	98	99	98	96	628	628	391

Notes: Initial Sample consists of individuals who completed the initial survey. Baseline Sample consists of individuals who completed the initial survey, who took the initial test, who joined Duolingo and who took the baseline belief survey. Columns (1)-(8) presents the averages. Column (9) tests the balance across the two waves, Column (10) tests the balance across the initial sample and the baseline sample, and Column (11) tests the balance across the four treatments of the baseline sample. Omitted ethnicity and major categories are other, omitted cohort category is masters or above, omitted Spanish category is Intermediate Spanish. 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 the question about their parents' highest level of education. Initial Test Score is the score in the initial Spanish test based on Duolingo lessons. Locus of Control variable is based on Locus of Control Survey (Rotter (1966)) and ranges between 0 and 8. A higher number signifies an external locus of control. Self Control variable is based on Self Control Survey (Tangney et al. (2004)) and ranges between 13 and 65. Growth Mindset is based on 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 Big Five Personality Trait Survey (Rammstedt and John (2007)) and the score ranges between 2 and 10 for each of them. Preference Survey controls are estimated using Convex Time Budget Sets as explained in Appendix Section A.7. Lower α signifies higher risk aversion. β below 1 indicates present-biasedness. Higher δ indicates patience and if δ is equal to 1, there is no discounting. Preference Survey Missing is equal to 1 if the student did not take the preference survey. Commitment is equal to 1 if the student states to be highly committed to learn Spanish and Interest is equal to 1 if the student states to be highly interested in learning Spanish. Serious Learner is a dummy variable which is equal to 1 if a student is learning Spanish for studying, living, working abroad purposes or for improving employment opportunities and 0 otherwise. 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).

Table B.2: Summary Statistics for the Analysis Sample

	Averages						P-Values	
			Treatments					
	(1) Baseline Sample	(2) Analysis Sample	(3) *No Information*	(4) *Average Information*	(5) *Low*	(6) *High*	(7) Baseline vs Analysis	(8) Joint Test across Treatments
Demographic Controls:								
Female	.71	.72	.76	.67	.73	.71	.038	.569
African-American	.05	.05	.07	.02	.05	.04	.039	.528
Asian	.44	.44	.33	.52	.46	.46	.713	.054
Caucasian	.34	.35	.45	.3	.28	.36	.68	.076
Hispanic	.08	.09	.1	.06	.12	.07	.571	.481
Age	23.01	22.93	23.86	23.01	22.51	22.36	.176	.136
Education Controls:								
Arts, Humanities, and Soc. Sci.	.25	.25	.23	.23	.28	.27	.593	.805
Business and Economics	.04	.04	.03	.04	.03	.07	.884	.662
Health and STEM	.63	.63	.64	.63	.61	.64	.532	.975
Freshman	.1	.1	.07	.04	.16	.14	.972	.018
Sophomore	.17	.18	.15	.2	.12	.24	.372	.145
Junior	.23	.23	.26	.29	.23	.16	.723	.216
Senior	.27	.26	.24	.22	.28	.3	.065	.58
GPA	3.42	3.43	3.46	3.42	3.42	3.39	.059	.789
SAT/ACT Math	.13	.12	.12	.14	.1	.14	.548	.943
SAT/ACT Reading	.09	.09	.12	.08	.09	.07	.653	.947
SAT/ACT Missing	.48	.48	.52	.46	.46	.47	.526	.804
Mother's Education	14.58	14.58	14.7	14	14.88	14.74	.894	.443
Father's Education	14.69	14.72	14.7	14.26	15.31	14.61	.631	.478
Initial Test Score	306.14	308.38	324.85	304.36	314.19	290.15	.206	.454
Personality Controls:								
Locus of Control	3.58	3.58	3.63	3.77	3.42	3.51	.928	.45
Growth Mindset	28.86	28.85	29.79	28.26	28.25	29.12	.806	.006
Self Control	44.51	44.41	44.52	44.81	44.39	43.92	.317	.906
Extroversion	6.22	6.23	6.2	6.27	6.3	6.14	.678	.94
Agreeableness	7.27	7.26	7.4	7.22	7.25	7.16	.478	.733
Neuroticism	5.99	6.04	6.02	5.87	6.23	6.03	.039	.644
Conscientiousness	7.39	7.38	7.34	7.35	7.34	7.5	.847	.868
Openness	7.37	7.42	7.46	7.41	7.06	7.73	.01	.044

Table B.2: Summary Statistics for the Analysis Sample

	Averages						P-Values	
			Treatments					
	(1) Baseline Sample	(2) Analysis Sample	(3) *No Information*	(4) *Average Information*	(5) *Low*	(6) *High*	(7) Baseline vs Analysis	(8) Joint Test across Treatments
Preference Survey Controls:								
α (Risk Aversion)	.57	.66	.67	.52	.74	.71	0	.249
β (Time Inconsistency)	1.04	1.04	1.05	1.05	1.03	1.05	.178	.927
δ (Patience)	.9991	.9992	1.002	.9981	.9979	.992	.716	.097
Preference Survey Missing	.09	.09	.1	.11	.09	.05	.015	.612
Other Controls:								
Commitment	.84	.84	.85	.86	.83	.84	.94	.931
Interest	.88	.88	.85	.89	.88	.88	.751	.807
Don't Know Spanish	.46	.46	.39	.43	.45	.55	.41	.139
Know Beginner Spanish	.48	.48	.54	.49	.48	.4	.796	.291
Know Elementary Spanish	.06	.06	.05	.09	.05	.04	.263	.656
Native English Speaker	.65	.66	.78	.59	.68	.61	.136	.021
Serious Learner	.52	.52	.5	.6	.45	.53	.488	.252
Income	97736	99332	91032	98670	96801	110869	.217	.676
Expenditure	284.14	279.72	264.4	255.59	303.76	295.38	.183	.507
Observations	391	371	92	94	93	92	391	371

Notes: Baseline Sample consists of individuals who completed the initial survey, who took the initial test, who joined Duolingo and who took the baseline belief survey. Analysis Sample is the subset of the Baseline Sample who completed Week 1 Belief Survey. Columns (1)-(6) presents the averages. Column (7) tests the balance across the baseline and the analysis samples and Column (8) tests the balance across the four treatments for the analysis sample. See notes of Appendix Table B.1 for detailed description of the variables and for the omitted categories.

Table B.3: Correlates of Baseline Beliefs

	(1) Baseline Beliefs About Others	(2) Baseline Beliefs About Self
Female	-0.459	-0.218
	(0.329)	(0.344)
African-American	2.087***	1.826**
	(0.793)	(0.840)
Asian	0.355	0.138
	(0.507)	(0.530)
Caucasian	-0.0811	-0.171
	(0.511)	(0.543)
Hispanic	0.0641	0.440
	(0.685)	(0.749)
Age	-0.0351	0.0332
	(0.0365)	(0.0452)
Arts, Humanities, and Soc. Sci.	-0.0153	-0.490
	(0.560)	(0.648)
Business and Economics	-0.330	-0.536
	(0.822)	(0.911)
Health and STEM	-0.0730	-0.549
	(0.546)	(0.628)
Freshman	0.410	0.897
	(0.659)	(0.680)
Sophomore	-0.814	-0.195
	(0.548)	(0.579)
Junior	-0.0299	-0.000570
	(0.491)	(0.467)
Senior	-0.176	0.0689
	(0.452)	(0.498)
GPA	-0.0371	0.155
	(0.409)	(0.397)
SAT/ACT Math	0.0394	-0.0556
	(0.353)	(0.364)
SAT/ACT Reading	-0.682**	-0.448
	(0.341)	(0.348)
Mother's Education	-0.00618	-0.0102
	(0.0436)	(0.0576)
Father's Education	0.0436	0.0306
	(0.0358)	(0.0409)
Initial Test Score	-0.00147	-0.00119
	(0.00122)	(0.00127)
Locus of Control	0.0724	0.108
	(0.0978)	(0.106)
Growth Mindset	0.0652*	0.0628
	(0.0382)	(0.0405)
Self Control	-0.0281	-0.0320
	(0.0225)	(0.0223)
Extroversion	-0.0123	0.0283
	(0.0935)	(0.0875)
Agreeableness	0.104	0.0838
	(0.106)	(0.109)
Neuroticism	0.140	0.0987
	(0.0924)	(0.0966)
Conscientiousness	0.0351	0.0164
	(0.133)	(0.132)
Openness	-0.0408	-0.0576
	(0.102)	(0.0995)
α (Risk Aversion)	-0.0728*	-0.0540
	(0.0396)	(0.0374)
β (Time Inconsistency)	1.315*	1.455*
	(0.719)	(0.771)
δ (Patience)	-8.215	-4.870
	(16.04)	(17.13)
Commitment	0.0303	0.108
	(0.436)	(0.420)
Interest	-0.501	-0.707
	(0.448)	(0.472)
Don't Know Spanish	0.665	0.0450
	(1.923)	(1.760)
Know Beginner Spanish	0.906	0.199
	(1.900)	(1.728)
Know Elementary Spanish	0.913	0.734
	(1.958)	(1.832)
Native English Speaker	0.550	0.259
	(0.415)	(0.406)
Serious Learner	0.0293	0.197
	(0.304)	(0.302)
Income	1.38e-06	9.71e-07
	(1.38e-06)	(1.34e-06)
Expenditure	0.000242	0.000617
	(0.000549)	(0.000513)
Wave	-0.470	-0.303
	(0.336)	(0.331)
Observations	391	391

Notes: The dependent variable in Column (1) is the baseline beliefs about the per-lesson improvement of the other participants and the dependent variable in Column (2) is the baseline beliefs about the per-lesson improvement of the self. Indicator variables for missing Preference Survey data and missing standardized test score data, and a constant are also included in the regressions. See notes of Appendix Table B.1 for detailed description of the variables and for the omitted categories. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table B.4: Baseline Beliefs about the Other Participants and the Self

	(1) *32-Lesson*	(2) *48-Lesson*	(3) *64-Lesson*	(4) *80-Lesson*	(5) *96-Lesson*	(6) *Per Lesson*
Other Participants	-1.939	-1.506	-2.900	-2.593	-3.882	-0.0420
	(4.402)	(4.859)	(5.367)	(5.361)	(5.589)	(0.0808)
Constant	137.3***	188.2***	241.4***	293.0***	345.7***	3.849***
	(6.657)	(7.751)	(8.850)	(9.814)	(10.69)	(0.142)
Observations	782	782	782	782	782	782
Subjects	391	391	391	391	391	391

Notes: Fixed Effect Regressions. Dependent variables in Columns (1)-(5) are the subjects' answers to the questions in Week 0 belief survey about the other participants and the self. For example, the dependent variable in Column (1) is the subjects' answer to the question of how much improvement in test scores results from completing 32 lessons in a month, on average. Column (6) depicts baseline beliefs for per-lesson improvement as defined in Equation 3. "Other participants" is an indicator variable that is equal to 1 if the belief question is about the other participants and 0 if the belief question is about the self. Robust standard errors in parentheses. Standard errors are clustered at the subject level. *** p<0.01, ** p<0.05, * p<0.1.

Table B.5: Effect of Treatments on Perceived Returns to Effort of Other Participants: Week 1 to Week 4

	(1)	(2)	(3)	(4)	(5)	(6)
	Week 4 Belief per Lesson-Week 1 Belief per Lesson					
Average	0.101	0.0969	0.0708	0.0763	0.0470	0.00549
	(0.232)	(0.231)	(0.244)	(0.253)	(0.262)	(0.277)
Low	-0.510*	-0.522*	-0.558**	-0.495*	-0.540*	-0.476*
	(0.274)	(0.275)	(0.279)	(0.282)	(0.285)	(0.286)
High	1.007***	0.994***	0.955***	0.912***	0.850***	0.890***
	(0.296)	(0.294)	(0.296)	(0.287)	(0.293)	(0.303)
Constant	0.185	-0.0599	0.948	0.528	11.80	11.74
	(0.251)	(0.434)	(0.993)	(1.625)	(10.45)	(11.02)
No-Information:						
Mean	-0.359	-0.378	-0.325	-0.547	-0.526	-0.479
Standard Deviation	0.166	0.170	0.182	0.198	0.212	0.269
P-values:						
Average=Low	0.022	0.022	0.021	0.045	0.042	0.106
Average=High	0.002	0.003	0.003	0.004	0.008	0.003
Low=High	0	0	0	0	0	0
Controls:						
Baseline Beliefs	Yes	Yes	Yes	Yes	Yes	Yes
Wave Fixed Effect	No	Yes	Yes	Yes	Yes	Yes
Demographic	No	No	Yes	Yes	Yes	Yes
Education	No	No	No	Yes	Yes	Yes
Personality & Preference	No	No	No	No	Yes	Yes
Other	No	No	No	No	No	Yes
Observations	334	334	334	334	334	334

Notes: OLS Regressions. Dependent variable is the perceived per lesson improvement of the other participants elicited in Week 1 subtracted from the perceived per lesson improvement of the other participants elicited in Week 4. Average is a dummy variable that is equal to 1 if a student is in the *Average* treatment and 0 otherwise. Similarly, Low (High) is a dummy variable that is equal to 1 if a student is in the *Low* (*High*) treatment and 0 otherwise. The omitted category is the *No Information* treatment. Baseline Beliefs are subjects' beliefs about the improvement of other participants for 5 different levels of effort (32, 48, 64, 80, and 96 lessons) elicited in Week 0. See notes of Appendix Table B.1 for detailed description of the control variables. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table B.6: Effect of Treatments on Perceived Returns to Effort of the Self: Week 0 to Week 1

	(1)	(2)	(3)	(4)	(5)	(6)
	Week 1 Belief per Lesson-Baseline Belief per Lesson					
Average	-0.856***	-0.861***	-0.918***	-0.911***	-0.870***	-0.945***
	(0.301)	(0.301)	(0.300)	(0.303)	(0.314)	(0.315)
Low	-1.101***	-1.090***	-1.135***	-1.188***	-1.157***	-1.200***
	(0.309)	(0.310)	(0.304)	(0.307)	(0.314)	(0.323)
High	0.488	0.498	0.445	0.427	0.467	0.388
	(0.310)	(0.309)	(0.308)	(0.308)	(0.321)	(0.331)
Constant	1.822***	2.278***	2.284**	3.796*	11.39	10.80
	(0.254)	(0.472)	(1.091)	(1.980)	(9.178)	(9.745)
No-Information:						
Mean	-0.459	-0.421	-0.388	-0.413	-0.444	-0.291
Standard Deviation	0.216	0.220	0.217	0.240	0.249	0.286
P-values:						
Average=Low	0.418	0.454	0.479	0.363	0.367	0.438
Average=High	0	0	0	0	0	0
Low=High	0	0	0	0	0	0
Controls:						
Baseline Beliefs	Yes	Yes	Yes	Yes	Yes	Yes
Wave Fixed Effect	No	Yes	Yes	Yes	Yes	Yes
Demographic	No	No	Yes	Yes	Yes	Yes
Education	No	No	No	Yes	Yes	Yes
Personality & Preference	No	No	No	No	Yes	Yes
Other	No	No	No	No	No	Yes
Observations	371	371	371	371	371	371

Notes: OLS Regressions.Dependent variable is the perceived per lesson improvement of the self elicited in Week 0 subtracted from the perceived per lesson improvement of the self elicited in Week 1. Average is a dummy variable that is equal to 1 if a student is in the *Average* treatment and 0 otherwise. Similarly, Low (High) is a dummy variable that is equal to 1 if a student is in the *Low* (*High*) treatment and 0 otherwise. The omitted category is the *No Information* treatment. Baseline Beliefs are subjects' beliefs about the improvement of the self for 5 different levels of effort (32, 48, 64, 80, and 96 lessons) elicited in Week 0. See notes of Appendix Table B.1 for detailed description of the control variables. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table B.7: Effect of Treatments on Perceived Returns to Effort of the Self: Week 1 to Week 4

	(1)	(2)	(3)	(4)	(5)	(6)
	Week 4 Belief per Lesson-Week 1 Belief per Lesson					
Average	0.175	0.170	0.136	0.102	0.0642	0.0676
	(0.279)	(0.278)	(0.281)	(0.282)	(0.286)	(0.288)
Low	-0.218	-0.229	-0.273	-0.193	-0.189	-0.0805
	(0.267)	(0.268)	(0.271)	(0.282)	(0.277)	(0.286)
High	1.170***	1.157***	1.122***	1.114***	1.140***	1.223***
	(0.279)	(0.280)	(0.284)	(0.285)	(0.294)	(0.306)
Constant	0.325	0.113	0.381	0.320	4.841	6.019
	(0.241)	(0.404)	(0.844)	(1.487)	(9.111)	(9.497)
No-Information:						
Mean	-0.497	-0.513	-0.458	-0.616	-0.596	-0.624
Standard Deviation	0.171	0.172	0.180	0.205	0.209	0.272
P-values:						
Average=Low	0.187	0.184	0.171	0.342	0.415	0.631
Average=High	0.001	0.001	0.001	0.001	0.001	0
Low=High	0	0	0	0	0	0
Controls:						
Baseline Beliefs	Yes	Yes	Yes	Yes	Yes	Yes
Wave Fixed Effect	No	Yes	Yes	Yes	Yes	Yes
Demographic	No	No	Yes	Yes	Yes	Yes
Education	No	No	No	Yes	Yes	Yes
Personality & Preference	No	No	No	No	Yes	Yes
Other	No	No	No	No	No	Yes
Observations	334	334	334	334	334	334

Notes: OLS Regressions. Dependent variable is the perceived per lesson improvement of the self elicited in Week 1 subtracted from the perceived per lesson improvement of the self elicited in Week 4. Average is a dummy variable that is equal to 1 if a student is in the *Average* treatment and 0 otherwise. Similarly, Low (High) is a dummy variable that is equal to 1 if a student is in the *Low* (*High*) treatment and 0 otherwise. The omitted category is the *No Information* treatment. Baseline Beliefs are subjects' beliefs about the improvement of the self for 5 different levels of effort (32, 48, 64, 80, and 96 lessons) elicited in Week 0. See notes of Appendix Table B.1 for detailed description of the control variables. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table B.8: Effect of Treatments on Perceived Returns to Effort of Others: Week 0 to Week 1

	(1)	(2)	(3)	(4)	(5)	(6)
	Week 1 Belief per Lesson-Baseline Belief per Lesson					
High	1.273**	1.253**	1.291**	1.029*	1.235**	1.000
	(0.592)	(0.591)	(0.593)	(0.597)	(0.605)	(0.632)
Info belongs to the Avg	-1.418***	-1.424***	-1.329***	-1.428***	-1.255**	-1.586***
	(0.469)	(0.473)	(0.500)	(0.470)	(0.521)	(0.560)
Info belongs to the Avg*High	0.779	0.814	0.823	0.901	0.494	0.817
	(0.713)	(0.725)	(0.744)	(0.769)	(0.771)	(0.791)
Constant	1.779***	2.092**	2.306	3.491	14.78	13.82
	(0.538)	(0.824)	(1.625)	(2.650)	(17.80)	(16.94)
Low:						
Mean	-0.692	-0.681	-0.760	-0.533	-0.588	-0.353
Standard Deviation	0.413	0.416	0.419	0.389	0.414	0.440
Controls:						
Baseline Beliefs	Yes	Yes	Yes	Yes	Yes	Yes
Wave Fixed Effect	No	Yes	Yes	Yes	Yes	Yes
Demographic	No	No	Yes	Yes	Yes	Yes
Education	No	No	No	Yes	Yes	Yes
Personality & Preference	No	No	No	No	Yes	Yes
Other	No	No	No	No	No	Yes
Observations	162	162	162	162	162	162

Notes: OLS Regressions. Dependent variable is the perceived per lesson improvement of the other participants elicited in Week 0 subtracted from the perceived per lesson improvement of the other participants elicited in Week 1. High is a dummy variable that is equal to 1 if a student is in the *High* treatment and 0 if the student is in the *Low* treatment. Info belongs to the Avg is a dummy variable that is equal to 1 if a student believes that the information is based on the average/median participant and 0 otherwise. The omitted category is the *Low* treatment. Baseline Beliefs are subjects' beliefs about the improvement of other participants for 5 different levels of effort (32, 48, 64, 80, and 96 lessons) elicited in Week 0. See notes of Appendix Table B.1 for detailed description of the control variables. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table B.9: Effect of Change in Beliefs on Final Spanish Assessment Test

	(1)	(2)	(3)	(4)	(5)	(6)
	Final Test Scores (Standardized)					
Change in Beliefs	0.0107	0.0131	0.0150	0.0168	0.0373	0.0379
	(0.0561)	(0.0551)	(0.0544)	(0.0501)	(0.0484)	(0.0488)
Initial Test Score (Std.)	0.650***	0.654***	0.648***	0.627***	0.646***	0.621***
	(0.0451)	(0.0447)	(0.0443)	(0.0434)	(0.0402)	(0.0502)
Constant	-0.0521	-0.522**	-0.872**	-1.454**	-7.423**	-6.596**
	(0.134)	(0.214)	(0.351)	(0.588)	(2.982)	(3.035)
Controls:						
Baseline Beliefs	Yes	Yes	Yes	Yes	Yes	Yes
Wave Fixed Effect	No	Yes	Yes	Yes	Yes	Yes
Demographic	No	No	Yes	Yes	Yes	Yes
Education	No	No	No	Yes	Yes	Yes
Personality & Preference	No	No	No	No	Yes	Yes
Other	No	No	No	No	No	Yes
Observations	328	328	328	328	328	328
First Stage F-Stat	12.38	12.35	12.55	13.88	13.42	13.36

Notes: 2SLS Regressions. Dependent variable is the standardized final test score. Change in beliefs is the perceived per lesson improvement of the other participants elicited in Week 0 subtracted from the perceived per lesson improvement of the other participants elicited in Week 1. The instruments are the treatment dummies. Initial Test Score (Std.) is the initial test score in standard deviation units. Baseline Beliefs are subjects' beliefs about the improvement of the other participants for 5 different levels of effort (32, 48, 64, 80, and 96 lessons) elicited in Week 0. See notes of Appendix Table B.1 for detailed description of the control variables. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table B.10: Effect of Change in Beliefs on Effort for Different Samples

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A:	Number of Lessons Completed					
Change in Beliefs	4.345**	5.642*	4.828**	5.486**	5.059**	4.734**
	(2.003)	(2.950)	(2.181)	(2.270)	(2.299)	(2.092)
Constant	2.461	-60.29	-16.00	-20.11	-23.36	-16.94
	(108.9)	(120.9)	(122.2)	(111.4)	(112.0)	(112.7)
Panel B:	Time Spent on Duolingo					
Change in Beliefs	17.94	35.76*	21.86	22.84*	25.44*	19.50
	(12.32)	(18.68)	(13.35)	(13.75)	(14.70)	(12.83)
Constant	-74.43	-347.2	-214.1	-134.8	-287.8	-176.4
	(568.2)	(684.8)	(644.1)	(588.8)	(580.5)	(580.0)
Panel C:	Drop Out (No Lessons Completed)					
Change in Beliefs	-0.0575***	-0.0811***	-0.0544***	-0.0608***	-0.0881***	-0.0665***
	(0.0196)	(0.0288)	(0.0208)	(0.0223)	(0.0243)	(0.0242)
Constant	0.717	1.587	0.680	0.886	1.214	2.599
	(1.333)	(1.490)	(1.473)	(1.382)	(1.534)	(1.528)
Controls:						
Baseline Beliefs	Yes	Yes	Yes	Yes	Yes	Yes
Wave Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Demographic	Yes	Yes	Yes	Yes	Yes	Yes
Education	Yes	Yes	Yes	Yes	Yes	Yes
Personality & Preference	Yes	Yes	Yes	Yes	Yes	Yes
Other	Yes	Yes	Yes	Yes	Yes	Yes
Observations	324	279	306	308	291	315
First Stage F-Stat	13.03	6.241	10.77	10.69	10.17	11.90
Sample	Final Survey Takers	No Other Tools	No Cheating in Tests	No Other Account	No Friends in the Study	Understood Instructions

Notes: 2SLS Regressions. Dependent variable in Panel A is the number of lessons completed during the four weeks of the experiment. Dependent variable in Panel B is minutes spent on Duolingo during the four weeks of the experiment. Dependent variable in Panel C is a dummy variable which is equal to 1 if a subject does not complete any Duolingo lessons during the study period (hence drops out from Duolingo) or 0 otherwise. Based on the answers reported in the final survey, Columns (1), (2), (3), (4), (5) and (6) restrict the sample to the individuals who took the final survey, who did not use any online or offline tools to learn Spanish (other than Duolingo), who did not use any sources such as internet, dictionary and/or asking somebody while taking the initial and final Spanish tests, who did not use another Duolingo account during the study period, who did not have friends participated in the study, and who correctly answered at least three questions (out of four questions) about the rules of the study, respectively. Change in beliefs is the perceived per lesson improvement of the other participants elicited in Week 0 subtracted from the perceived per lesson improvement of the other participants elicited in Week 1. The instruments are the treatment dummies. Baseline Beliefs are subjects' beliefs about the improvement of other participants for 5 different levels of effort (32, 48, 64, 80, and 96 lessons) elicited in Week 0. See notes of Appendix Table B.1 for detailed description of the control variables. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table B.11: Effect of Change in Beliefs on Effort-Min

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A:	Number of Lessons Completed					
Change in Beliefs	4.096*	4.072*	3.719*	3.638**	3.839**	3.866**
	(2.115)	(2.111)	(1.996)	(1.847)	(1.925)	(1.875)
Constant	26.90***	21.63***	-20.53	-44.32**	-8.841	9.833
	(3.449)	(6.399)	(13.65)	(21.76)	(106.4)	(106.2)
Panel B:	Time Spent on Duolingo					
Change in Beliefs	17.91	17.81	15.97	15.72	16.37	16.33
	(11.61)	(11.60)	(11.12)	(10.64)	(11.55)	(11.36)
Constant	134.6***	124.6***	-115.7	-188.0	-70.46	-51.22
	(18.48)	(38.03)	(89.18)	(145.7)	(546.1)	(557.1)
Panel C:	Drop Out (No Lessons Completed)					
Change in Beliefs	-0.0359*	-0.0362*	-0.0353*	-0.0307	-0.0337*	-0.0344*
	(0.0213)	(0.0213)	(0.0207)	(0.0193)	(0.0202)	(0.0204)
Constant	0.136***	0.114*	0.289**	0.316	0.615	0.660
	(0.0366)	(0.0687)	(0.121)	(0.211)	(1.430)	(1.384)
Controls:						
Baseline Beliefs	Yes	Yes	Yes	Yes	Yes	Yes
Wave Fixed Effect	No	Yes	Yes	Yes	Yes	Yes
Demographic	No	No	Yes	Yes	Yes	Yes
Education	No	No	No	Yes	Yes	Yes
Personality & Preference	No	No	No	No	Yes	Yes
Other	No	No	No	No	No	Yes
Observations	391	391	391	391	391	391
First Stage F-Stat	6.172	6.172	6.432	7.037	6.247	6.312

Notes: 2SLS Regressions. Dependent variable in Panel A is the number of lessons completed during the four weeks of the experiment. Dependent variable in Panel B is minutes spent on Duolingo during the four weeks of the experiment. Dependent variable in Panel C is a dummy variable which is equal to 1 if a subject does not complete any Duolingo lessons during the study period (hence drops out from Duolingo) or 0 otherwise. Change in beliefs is the perceived per lesson improvement of the other participants elicited in Week 0 subtracted from the perceived per lesson improvement of the other participants elicited in Week 1. The instruments are the treatment dummies. Baseline Beliefs are subjects' beliefs about the improvement of other participants for 5 different levels of effort (32, 48, 64, 80, and 96 lessons) elicited in Week 0. See notes of Appendix Table B.1 for detailed description of the control variables. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table B.12: Effect of Change in Beliefs on Effort-Max

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A:			Number of Lessons Completed			
Change in Beliefs	4.007*	4.000*	3.437	3.353	3.383*	3.298*
	(2.304)	(2.302)	(2.208)	(2.109)	(2.000)	(1.995)
Constant	21.36***	18.68**	-21.90	-48.54**	-38.70	-31.94
	(6.128)	(7.776)	(15.59)	(24.31)	(108.7)	(109.6)
Panel B:			Time Spent on Duolingo			
Change in Beliefs	17.07	17.06	13.99	14.29	13.71	12.97
	(12.35)	(12.36)	(12.02)	(11.77)	(11.57)	(11.57)
Constant	111.4***	112.8**	-119.6	-205.6	-190.0	-215.0
	(32.07)	(44.65)	(96.96)	(159.9)	(544.1)	(562.4)
Panel C:			Drop Out (No Lessons Completed)			
Change in Beliefs	-0.0376*	-0.0378*	-0.0366*	-0.0323	-0.0341*	-0.0362*
	(0.0218)	(0.0218)	(0.0216)	(0.0209)	(0.0199)	(0.0207)
Constant	0.190***	0.146*	0.312**	0.364	0.925	1.122
	(0.0598)	(0.0834)	(0.141)	(0.226)	(1.388)	(1.336)
Controls:						
Baseline Beliefs	Yes	Yes	Yes	Yes	Yes	Yes
Wave Fixed Effect	No	Yes	Yes	Yes	Yes	Yes
Demographic	No	No	Yes	Yes	Yes	Yes
Education	No	No	No	Yes	Yes	Yes
Personality & Preference	No	No	No	No	Yes	Yes
Other	No	No	No	No	No	Yes
Observations	391	391	391	391	391	391
First Stage F-Stat	6.558	6.566	6.609	6.470	7.473	6.985

Notes: 2SLS Regressions. Dependent variable in Panel A is the number of lessons completed during the four weeks of the experiment. Dependent variable in Panel B is minutes spent on Duolingo during the four weeks of the experiment. Dependent variable in Panel C is a dummy variable which is equal to 1 if a subject does not complete any Duolingo lessons during the study period (hence drops out from Duolingo) or 0 otherwise. Change in beliefs is the perceived per lesson improvement of the other participants elicited in Week 0 subtracted from the perceived per lesson improvement of the other participants elicited in Week 1. The instruments are the treatment dummies. Baseline Beliefs are subjects' beliefs about the improvement of other participants for 5 different levels of effort (32, 48, 64, 80, and 96 lessons) elicited in Week 0. See notes of Appendix Table B.1 for detailed description of the control variables. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table B.13: Effect of Change in Beliefs about the Self on Effort

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A:	Number of Lessons Completed					
Change in Beliefs	4.238*	4.219*	3.370	3.261	3.574*	3.392
	(2.318)	(2.314)	(2.177)	(2.103)	(2.166)	(2.113)
Constant	22.28***	17.31***	-26.38*	-50.99**	-7.295	20.13
	(4.592)	(6.664)	(14.73)	(22.78)	(104.3)	(102.7)
Panel B:	Time Spent on Duolingo					
Change in Beliefs	18.71	18.67	14.27	13.45	15.03	13.63
	(13.45)	(13.43)	(12.61)	(12.34)	(12.65)	(12.54)
Constant	110.8***	102.4**	-139.3**	-214.9*	-40.42	9.731
	(26.45)	(42.63)	(70.36)	(124.4)	(776.4)	(784.2)
Panel C:	Drop Out (No Lessons Completed)					
Change in Beliefs	-0.0369	-0.0371	-0.0348	-0.0328	-0.0405*	-0.0427*
	(0.0230)	(0.0230)	(0.0220)	(0.0219)	(0.0225)	(0.0228)
Constant	0.158***	0.133*	0.325***	0.381*	0.980	0.942
	(0.0480)	(0.0725)	(0.123)	(0.211)	(1.380)	(1.365)
Controls:						
Baseline Beliefs	Yes	Yes	Yes	Yes	Yes	Yes
Wave Fixed Effect	No	Yes	Yes	Yes	Yes	Yes
Demographic	No	No	Yes	Yes	Yes	Yes
Education	No	No	No	Yes	Yes	Yes
Personality & Preference	No	No	No	No	Yes	Yes
Other	No	No	No	No	No	Yes
Observations	371	371	371	371	371	371
First Stage F-Stat	11.24	11.30	11.74	12.54	12.01	12.05

Notes: 2SLS Regressions. Dependent variable in Panel A is the number of lessons completed during the four weeks of the experiment. Dependent variable in Panel B is minutes spent on Duolingo during the four weeks of the experiment. Dependent variable in Panel C is a dummy variable which is equal to 1 if a subject does not complete any Duolingo lessons during the study period (hence drops out from Duolingo) or 0 otherwise. Change in beliefs is the perceived per lesson improvement of the self elicited in Week 0 subtracted from the perceived per lesson improvement of the self elicited in Week 1. The instruments are the treatment dummies. Baseline Beliefs are subjects' beliefs about the improvement of the self for 5 different levels of effort (32, 48, 64, 80, and 96 lessons) elicited in Week 0. See notes of Appendix Table B.1 for detailed description of the control variables. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.