Learning Spanish with Babbel: Oral Proficiency Outcomes in App-Based Foreign Language Learning

Report on the findings of an efficacy study conducted by

Dr. Shawn Loewen
Professor, Second Language Studies Program
Michigan State University

Daniel R. Isbell
PhD Candidate, Second Language Studies
Michigan State University

&

Zachary Sporn
Senior Communications Manager: Research and Pedagogy
Babbel

Executive Summary

Babbel

Eighty-five participants were recruited at Michigan State University to learn Spanish with the language learning application Babbel over a period of approximately three months. Fifty-four participants persisted in Spanish study and successfully completed all study procedures. In general, participants made statistically significant gains in Spanish oral proficiency, grammar knowledge, and vocabulary knowledge; these gains were predicted by the amount of participants' total Babbel study hours. The results provide the first evidence that learning with Babbel leads to improved ability to communicate orally in Spanish.

Key Findings:

- Virtually all study participants made a measurable gain in their grammar and vocabulary knowledge and/or ability to communicate orally in Spanish. Learning gains in terms of oral proficiency, grammar knowledge, and vocabulary knowledge were all associated with how much time participants spent using Babbel.
- The findings show that Babbel's pedagogical approach enables learners to transfer receptive, input-based learning and explicit grammar and vocabulary instruction to communicative (oral) production at the Novice and Intermediate levels.
- On a whole, learners in this study increased their oral proficiency, as measured by a one or more sublevel (median= 1.0; mean = .70) improvement on the American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview-computer version (OPIc).
- The OPIc is a web-based assessment, delivered via a computer, which simulates a live interview. ACTFL OPIc ratings provide a metric for describing spoken functional ability in a foreign language. Ratings are divided into the following levels: Novice, Intermediate, Advanced and Superior. ACTFL levels can be further broken down into sublevels (e.g., Novice Low, Novice Mid and Novice High) for more granularity in describing language proficiency.
- 70% of those participants who studied a minimum of six hours over the duration of the study improved their oral proficiency by at least one ACTFL sublevel.
- 78% percent of those participants who studied at least 15 hours over the course of the study improved by at least one ACTFL sublevel.
- 11% of the total sample improved by two or more ACTFL sublevels.
- 32 participants (61% of the total sample) were rated Novice Low on their initial OPIc test. Of the participants who started at Novice Low or below, more than half improved by at least one sublevel with 12.4 hours of study on average.
- Of the 21 learners who started at novice-mid and above, 71% improved, after studying an average of 13.9 hours.
- Nearly all participants improved their Spanish grammar and vocabulary knowledge.
- The Spanish grammar and vocabulary tests used in this project were adapted from tests used in previously published research to reflect Babbel's lesson content, as well as to fairly assess participants' initially low levels of Spanish knowledge. Because the grammar and vocabulary tests were linked to Babbel's lesson content, they represent a form of achievement test. Gains made on these tests establish participants' ability to recall and apply what they had learned with Babbel.

Learning Spanish with Babbel

Introduction

Babbel is an online language learning platform and the world's highest grossing language learning mobile application. Babbel has over one million paying subscribers who study one (or more) of 14 different languages. Spanish was selected as the target language for this study, as it is a popular second/foreign language (hereafter "L2") to learn in the United States, where the research team at Michigan State University are located, as well as globally. Spanish is also one of Babbel's most-developed offerings, with a wide selection of courses from novice to intermediate. Babbel's Spanish courses for English native speakers include the *Spanish Beginner's Courses 1-6, Spanish for your Vacation, Conversations at Work,* and numerous thematically organized courses focusing on grammar, pronunciation and specialized vocabulary.

Babbel features bite-sized, contextualized lessons created by a team of 150 linguists, language teachers and instructional designers. Most lessons feature audio dialogues recorded by native speakers, and a variety of exercise types, encompassing speaking, writing, reading and listening skills. Among other tools and resources, Babbel also features a vocabulary "Review Manager" based on the empirically proven concept of spaced repetition. New words that learners encounter in Babbel lessons are automatically added to their personalized Review Manager, so they can be revised and consolidated at increasing time intervals until the word is mastered.

Using Babbel has already been shown by at least one study to considerably improve users' receptive second-language (L2) knowledge (Vesselinov & Grego, 2016). What is still less clear is to what degree improvements in productive and/or communicative language abilities (e.g., oral communication) might improve after using Babbel. Indeed, considerable doubt has been expressed in the L2 and computer assisted language learning (CALL) literature as to whether meaningful development of oral language abilities is possible in the context of commercial online language learning platforms (e.g., Lord, 2015). Thus, the present study expands on previous efforts to examine Babbel's effect on L2 development by including a widely used measure of oral L2 proficiency, the American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview-computer version (OPIc). Additionally, the study measured discrete, receptive linguistic knowledge of vocabulary, and receptive and productive knowledge of grammar. Finally, in order to better understand the role of motivational factors and the experiences of individual participants, the research team at Michigan State University collected questionnaire and interview data.

Research Questions

The following research questions were addressed:

1. What linguistic gains are made by English native speaker university students who exclusively use Babbel for autonomous study of Spanish as a second language (L2)? Both pre- and post-tests consist of:
    a. an OPIc test
    b. a vocabulary test
    c. a grammar test
2. Is there a relationship between how learners use Babbel and gains, if any, on test scores?

A forthcoming publication in a peer-reviewed linguistics journal will address several additional research questions which are outside the scope of this report.

Methods

The researchers employed a within-subjects quasi-experimental design with pretests and posttests to examine oral proficiency, grammar, and vocabulary learning outcomes after a roughly three-month period of using Babbel. To examine motivational factors during the course of the study, researchers took additional repeated measures of learners' interest in Spanish and desire to use Babbel. Further methodological details are included in the following subsections.

Participant recruitment and interviews, distribution of surveys, and analysis of test and survey results were conducted independently by the research team at Michigan State University. Participants accessed the Oral Proficiency Interview-computer version (OPIc) test via computer terminals at Michigan State University; these tests were supervised by the research team. Babbel provided the research team at MSU with financial and technical support, including the cost of OPIc testing. Every official OPIc was scored by two raters trained and certified by the American Council on the Teaching of Foreign Languages (ACTFL). Babbel's Analytics team provided the researchers detailed overview of participants' learning activity each week for the duration of the study. These data metrics included (but were not limited to) the number of minutes spent learning on Babbel, the number and names of Babbel Spanish lessons completed, and how often participants logged into the app weekly.

Participants

Participants were recruited at Michigan State University. Recruiting efforts included researcher visits to classes related to language learning (e.g., TESOL, linguistics) and campus-wide posting of recruitment flyers. Additionally, "snowball" recruiting was encouraged; participants were asked to pass along researchers' contact information to interested roommates, friends, family, and/or coworkers.

In total, 85 participants began the study. Two participants requested to withdraw from the study after beginning and their data are not considered here. Among the 83 remaining participants, 58 were considered eligible for post-testing by meeting at least two of the following three criteria:

1) at least 3 hours of total study,

2) a weekly average of at least 20 minutes of study (participants varied in the total number of weeks they were in the study; see Procedures section),
3) no period of 4 or more consecutive weeks of not using Babbel (i.e., people who had effectively quit studying).

51 participants met all three criteria and 7 met two. Of the 58 participants eligible for post-testing, 54 of them came in for post-testing appointments and completed all study procedures. This constitutes a study attrition rate of 36% (54/85); the attrition rate for Babbel use was 32% (58/85).

The final sample of 54 participants had an average age of approximately 24 years (median = 22 years). On average they had previously completed approximately two classroom-based Spanish courses, mostly in high school. Participants' average self-rated Spanish oral proficiency was between Novice Low and Novice Mid on the ACTFL scale (see Instruments subheader for more details on the ACTFL scale).

Instruments

Details on the surveys and tests used in the study follow:

The ACTFL Oral Proficiency Interview – Computer (OPIc) is a standardized measure of functional oral proficiency published by the American Council on the Teaching of Foreign Languages (ACTFL). The OPIc requires examinees to orally respond to between 12 and 17 speaking prompts, depending on the level of the test, which are spoken aloud by a virtual avatar on a computer screen in front of them. The test thus elicits evidence of conversational and interpersonal language ability. OPIc responses are scored by two certified ACTFL raters. Results are reported according to ACTFL's model of language proficiency, the ACTFL Guidelines 2012. The Guidelines feature descriptors of proficiency at 11 different (sub)levels ranging from Novice Low to Distinguished. These descriptors give readily-understandable meaning to the OPIc scores. Descriptor excerpts from relevant levels are presented in Table 1 below; please refer to the Guidelines for full versions (available here). OPIc test-takers who cannot produce any of the target language, (in this case Spanish) are designated as *Unratable* (UR). For our quantitative analyses, we converted ACTFL ratings to integers, with Unratable corresponding to 0 and Intermediate High corresponding to 6. (This numeric conversion is also used in Isbell, Winke, & Gass (in press) and Thompson, Cox, & Knapp (2016), etc.).

Table 1.
ACTFL Guidelines – with excerpts from the ACTFL

Level (Abbreviation)	Descriptor Excerpt

Unratable	Responses that are completely in English or off-topic are considered unratable.
Novice Low (NL)	"no real functional ability… cannot therefore participate in a true conversational exchange"
Novice Mid (NM)	"communicate minimally by using a number of isolated words and memorized phrases… may say only two or three words at a time"
Novice High (NH)	"conversation is restricted to a few of the predictable topics necessary for survival… language consists primarily of short and sometimes incomplete sentences"
Intermediate Low (IL)	"able to handle successfully a limited number of uncomplicated communicative tasks… Conversation is restricted to some of the concrete exchanges and predictable topics necessary for survival"
Intermediate Mid (IM)	"able to handle successfully a variety of uncomplicated communicative tasks in straightforward social situations… related to self, family, home, daily activities… [etc.]"
Intermediate High (IH)	"able to handle successfully uncomplicated tasks and social situations requiring an exchange of basic information related to their work, school, … and areas of competence."

Note: All excerpts from ACTFL (2012, pp. 7-9).

ACTFL provides different forms of the OPIc to assess different proficiency levels. For the current study, the researchers primarily used "Form 2" of the ACTFL OPIc, which targets the ability range of Novice Mid to Intermediate Mid. Scores from Novice Low to Intermediate High may be awarded on the basis of Form 2 performance. Although the difficulty level of the form was pre-selected, participants filled out a background questionnaire on the OPIc that was used to automatically select relevant speaking prompts that populated the form each participant received. A subset of seven participants who received ratings of Intermediate Low to Intermediate High on the pretest, were assigned Form 3 of the OPIc (which has as score range of Novice High to Advanced Low) on the posttest to ensure that any learning gains were detectable. The Spanish version of the OPIc has high reliability: Inter-rater reliability was reported as $\rho = .94$, absolute rater agreement as 77%, and absolute/adjacent agreement (exact or within +/- 1 sublevel) 96% for over 8,000 Spanish OPIc exams scored between 2012-2014 (Cubbellotti, 2015).

Grammar test. Researchers used a grammar test modeled on the grammar test in Leonard and Shea (2017) which contained 30 error identification and correction items. The grammar test items were modified to reflect structures introduced throughout the various Babbel Spanish courses with assistance from Babbel's Spanish course designers. Each item contained a sentence with one grammatical error (e.g., verb agreement) and was scored on a 2-point scale: 1 point for correctly identifying the error, and 1 point for providing an appropriate correction. The reliability (Cronbach's alpha) of the grammar test was .93 ($n = 83$) at pretest and .92 ($n = 54$) at posttest.

Vocabulary test. Researchers used a vocabulary test based on the specifications of the LexTALE-Esp (Izura, Cuetos, & Brysbaert, 2014), which is used to differentiate different

levels of vocabulary knowledge.. Like the LexTALE-Esp, this test featured a list of 90 words, with 60 real and 30 pseudo words, and a simple yes/no response format (i.e., participants put a check mark by Spanish words they recognized). However, to better capture the growth of Spanish vocabulary in a low-proficiency and relatively low-exposure instructed environment like a mobile language app, many of the low frequency real word items were substituted with vocabulary items found throughout the Babbel Spanish curriculum. Thus, this version of LexTALE-Esp reflects learning achievement rather than generalizable vocabulary size.

Researchers followed Izura et al.'s (2014) suggested scoring method: indicating a real word earned 1 point, indicating a non-word earned -2 points, and all unmarked words were awarded 0 points. Reliability (Cronbach's alpha) was .86 for the pretest (n = 83) and .87 for the posttest (n = 54).

Background questionnaire. The background questionnaire contained questions about participants' linguistic background, previous experiences with Babbel and other language learning apps, Spanish learning experience, and motivation to learn Spanish and use an app to learn a language.

Progress survey. The progress survey contained the same questions about motivation from the background questionnaire. It contained additional questions related to Babbel enjoyment, perceptions of learning efficacy, communication with other study participants, and time spent per week communicating in Spanish with another learner or speaker. It also provided space for participants to leave open-ended feedback on their experiences studying Spanish on Babbel.

Post-study survey. The post-study survey mirrors the content of the progress survey described previously. Additionally, it elicited percentages of mobile and desktop use, likelihood to continue studying Spanish, likelihood to continue studying Spanish on Babbel, and likelihood to study another language on Babbel.

Procedures

Study procedures can be divided into three main phases: Pre-testing, Babbel study, and Post-testing. The total duration of study procedures varied somewhat across participants, due to staggered Pre-testing and Post-testing; the average study duration was 84 days (12 weeks).

Pre-testing. Over the course of two weeks, all participants came to a pre-testing appointment in a lab equipped with computers and headsets. One of the researchers oversaw all pre-testing procedures and provided technical support. Upon arrival, participants filled out a screening form to verify that they (a) grew up speaking English (alongside any other languages) and (b) that they did not have high levels of Spanish proficiency. Participants self-rated their Spanish proficiency based on the overall oral communication descriptors from the ACTFL Guidelines 2012. No participant self-rated above the Intermediate Mid level. After screening, participants received information on study procedures and gave their informed consent to participate.

Participants then completed the background questionnaire, vocabulary test, grammar test, and ACTFL OPIc, in that order. All participants received a code providing free access to all Spanish learning content on Babbel's app and web browser versions for the duration of the study.

Babbel study. Participants studied Spanish on Babbel for a period of roughly 12 weeks, varying slightly for each participant based on their pre-testing and post-testing appointment. Participants were asked to study for a minimum of 15 minutes per day on average and were assured that occasionally taking a day off would not disqualify them from the study. Participants were also encouraged to study more than 15 minutes per day if they wanted. Participants were free to study whichever lessons they liked. They could also decide whether or not to revise vocabulary using Babbel's Review Manager.

Each week, participants received emails from Babbel and from the researchers. Emails from Babbel were upbeat reminders to keep up with studying Spanish. Emails from the researchers contained week-by-week data on how much time participants spent on Babbel. Additionally, the researchers sent out progress surveys at roughly week 4 and week 8 of the study.

Post-testing. Post-testing procedures were similar to pre-testing procedures. Instead of starting with the background questionnaire, participants took the vocabulary, grammar, and OPIc tests (in that order) before completing the post-study survey. After completing these tasks, participants were compensated with a $75 gift card and a 1-year subscription to Babbel.

Analyses

Quantitative analyses primarily entailed linear mixed-effects regression models. Conceptually similar to repeated measures ANOVA, linear mixed-effects regressions allow for observations to be correlated within-subjects, thereby accounting for the relationship between a person's initial and subsequent performance. Mixed-effects regressions are more flexible than repeated measures ANOVA. For instance, continuous predictor variables are easily accommodated, and more detailed random effects can be specified to account for structure in the data.

Results

The quantitative learning results for the 54 learners in the final sample are presented in the immediately following subsections, including descriptive statistics and linear mixed-effects regression models.

Descriptive Statistics and Correlations

Basic descriptive statistics for key study variables of interest are presented in Table 2. At the top of the table are the linguistic outcomes, including pretest, posttest, and change scores. Towards the bottom of the table are variables we predicted would influence the linguistic outcomes, including interest in learning Spanish, interest in using Babbel, hours spent using Babbel during the study, and classroom learning experience prior to the study.

Table 2.
Descriptive statistics for key study variables

	n	mean	SD	median	min	max
OPIc Pretest	54	1.81	1.33	NM	UR	IH
OPIc Posttest	54	2.52	1.50	NH	NL	IH
OPIc Change	54	0.70	0.74	1	-1	3
Grammar Pretest	54	11.06	11.79	7	0	45
Grammar Posttest	54	20.24	13.63	18	0	48
Grammar Change	54	9.19	7.67	8	-5	28
Vocabulary Pretest	54	12.89	10.04	10	1	35
Vocabulary Posttest	54	19.63	11.81	21	1	43
Vocabulary Change	54	6.74	6.66	6	-5	22
Interest in Learning Spanish* (-3 to 3)	54	2.01	0.86	2.08	-0.67	3.00
Interest in Using Babbel* (-3 to 3)	54	1.49	1.02	1.84	-1.17	3.00
Babbel Study Hours	54	11.61	7.27	9.75	2.27	27.75
Prior Classroom Learning**	52	2.08	2.06	2	0	10

*Computed by integrating across (up to) 4 time points. Most (38) participants reported their motivation at 4 time points; 11 had only 3 observations and 5 had only 2 observations. All participants had pretest and posttest data. **The sum of secondary years of and post-secondary semesters of classroom Spanish courses.

On a whole, learners in this study increased their oral proficiency by one sublevel on the ACTFL scale (median= 1.0; mean = .70). They also increased their grammar scores by 9.19 points and vocabulary scores by 6.74 points on average. Learners spent an average of 11.61 hours learning Spanish on Babbel over the approximately three-month duration of the study, or less than an hour per week. Prior to the study, participants had taken roughly two Spanish courses in high school or university, though several participants had no prior formal instruction in the language.

To begin understanding the relationships among these variables of interest, it is informative to examine bivariate correlations (Table 3). The three types of linguistic gains were moderately correlated, with grammar and vocabulary gains having a somewhat stronger correlation (r = .65). Gains in OPIc scores had moderate associations with Spanish interest, Babbel interest, and time spent learning on Babbel. In comparison, vocabulary and grammar gains had stronger associations with time spent on Babbel and somewhat weaker associations with the interest variables. Prior classroom Spanish learning had small to no relationship with other study variables except for interest in learning Spanish.

Table 3.
Correlations among study variables

	1	2	3	4	5	6
1. OPIc Change	1.00					
2. Grammar Change	.54	1.00				
3. Vocabulary Change	.48	.65	1.00			
4. Spanish Interest	.29	.07	.21	1.00		

5. Babbel Interest	.31	.22	.17	.56	1.00	
6. Babbel Study Hours	.30	.49	.49	.22	.36	1.00
7. Prior Classroom Learning	.12	-.01	-.11	.30	.16	.06

In-depth: Changes in OPIc Scores. Changes in OPIc scores are presented in greater depth in the following contingency table (Table 4). As Table 4 on the following page shows, most participants started the study at Novice Low or Novice Mid proficiency; by the end of the study the majority of participants were at Novice Mid. All participants within the shaded boxes made gains of at least one sublevel.

Table 4.
Contingency table of pre- and posttest OPIc scores. The number in each box represents the number of participants whose scores at pre- and post-test fell within each sublevel (Total N=54). All those within the shaded boxes increased in their OPIc score from pre- to post-test.

		Posttest						
		UR	NL	NM	NH	IL	IM	IH
Pretest	UR	0	1	0	0	0	0	0
	NL	0	16	12	3	1	0	0
	NM	0	0	2	6	1	0	0
	NH	0	0	1	0	3	1	0
	IL	0	0	0	0	1	2	0
	IM	0	0	0	0	0	1	2
	IH	0	0	0	0	0	0	1

Table 5 summarizes the varying magnitudes of oral proficiency gains. The most common outcome was a gain of 1 ACTFL sublevel (n = 26; 48%). Five participants demonstrated a substantial gain of 2 sublevels, and one participant made an outstanding increase of 3 sublevels, going from Novice Low to Intermediate Low. One participant

showed a net decrease in oral proficiency, possibly due to an uncharacteristically poor performance at posttest or differences in rater severity at pretest and posttest.

Table 5.
Net change in OPIc score from pretest to posttest

Change	n	%
-1	1	2
0	21	39
+1	26	48
+2	5	9
+3	1	2

Mixed-Effect Regression Analyses

Mixed-effect regression analyses allow for the relationships between learning outcomes and key variables of interest to be modeled and tested statistically. Separate linear mixed-effects regression (LMER) models for each learning outcome (oral proficiency, grammar, and vocabulary) were arrived through an iterative model selection process and final models are summarized following.

Oral Proficiency (OPIc Scores). The LMER model for oral proficiency is summarized in Table 6. Significant fixed-effect interactions of Time (pretest and posttest) with Babbel study hours and Time with interest in learning Spanish were obtained. In other words, it can be said that increases in OPIc scores over time were dependent on the amount of time a participant spent on Babbel and also on a participant's overall level of interest in learning Spanish. According to model predictions, participants with ambivalence toward learning Spanish (i.e., 0 overall motivation) would need roughly 33 hours of Babbel study to achieve an increase of one sublevel on the OPIc. However, no participant logged that many hours, and each level of interest in learning Spanish accounted for an increase of nearly a third of a sublevel from pretest to posttest, supplementing the Babbel study time. For example, participants with at least some interest in learning Spanish (i.e., maintaining an overall motivation score of 1 on the surveys) would be predicted to require only approximately 23 hours of Babbel study to increase their ACTFL proficiency level.

Additionally, prior classroom learning experience had a statistically significant and considerable effect on oral proficiency at the pretest. Roughly three Spanish courses (including courses taken several years ago for many participants) predicted an initial proficiency rating of one sublevel. Interestingly, no significant interaction effect between classroom experience and time was obtained, which can be interpreted as there being no effect (beneficial or detrimental) of previous classroom instruction on oral proficiency development when using Babbel. In other words, oral proficiency increases were not just a case of experienced learners recovering atrophied skills (though that possibility cannot be ruled out entirely on an individual basis).

Table 6.
Summary of OPIc score linear mixed-effects regression model.

Fixed Effects	B (SE)	β	p
Intercept	1.17 (0.23)		
Time	-0.10 (0.26)	-0.03	0.70
Classroom Experience	0.31 (0.08)	0.42	<.001
Time x Babbel Study Hours	0.03 (0.01)	0.16	0.03
Time x Spanish Interest	0.29 (0.11)	0.19	0.04

Random Effects	Variance (SD)
Participant (intercept)	1.23 (1.11)
Participant: Time (slope)	0.11 (0.33)

Model $R^2_{marginal}$ = .29, $R^2_{conditional}$ = .91. χ^2 vs. unconditional model = 43.217, df = 4, p <.001.

Figure 1 is a visualization of the effect of study time on oral proficiency development. Each panel corresponds to different levels of Babbel study time; participants with relatively little study time are featured in the leftmost panel and those with the most study time in the rightmost panel. The blue line with shaded standard error shows the LMER model-predicted pretest and posttest proficiency levels, and the grey lines correspond to the observed ratings for each individual participant (with some jittering to prevent over-plotting). The blue lines become slightly and progressively steeper as total Babbel study time increases, highlighting the significant interaction reported previously. It is also worth noting the considerable individual variation in initial proficiency and learning gains. As indicated by the random effects in Table 6, initial proficiency level had a standard deviation larger than 1 sublevel. The change in proficiency over time also varied across participants, with a standard deviation of .33 sublevels. The individual lines in Figure 1 reflect this, with some participants starting with higher or lower levels of proficiency. Similarly, some participants' lines are flat while others are quite steep.

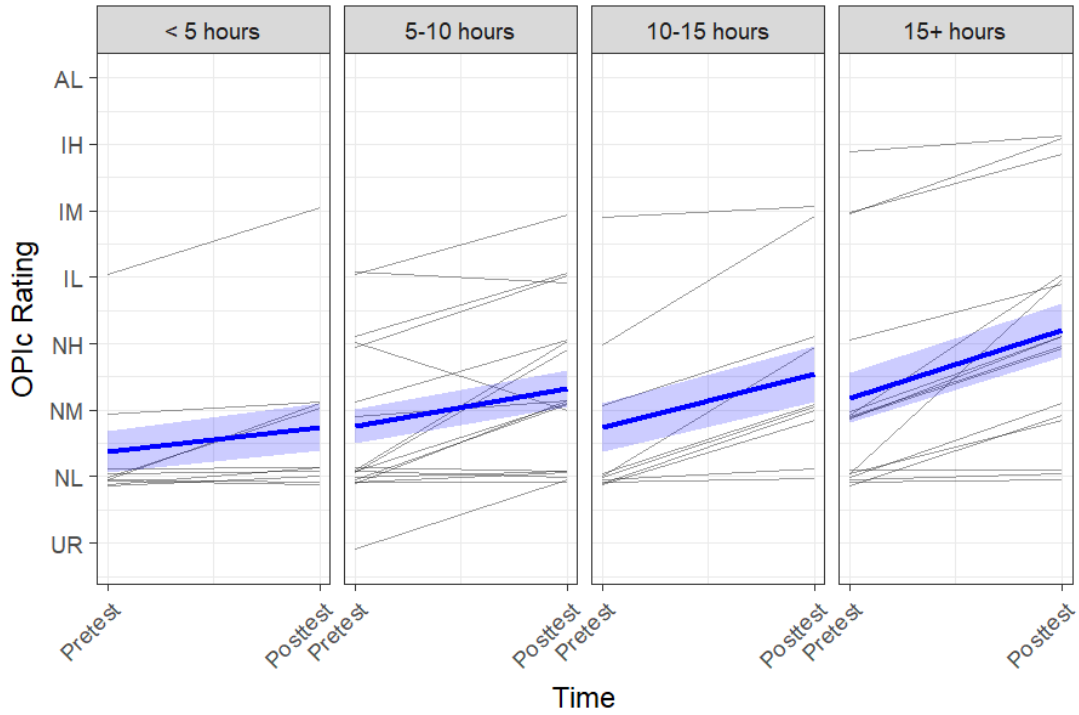


Figure 1. Changes in oral proficiency over time according to number of hours studied on Babbel. Participants were divided into four groups of Babbel study hours to illustrate trends. Blue lines and shaded areas indicate model-predicted averages and standard errors, respectively. Grey lines are based on individual participants' observed test scores.

Contextualizing Participants' Oral Proficiency Gains

The ACTFL OPIc speaking test assigns a rating based on the ACTFL Proficiency Guidelines. These guidelines describe "what learners can do with language in terms of speaking, writing, listening, and reading in real-world situations in a spontaneous and non-rehearsed context." (ACTFL Proficiency Guidelines, 2012).

For the speaking skill, these guidelines identify five major levels of proficiency: Distinguished, Superior, Advanced, Intermediate, and Novice, which are further subdivided into High, Mid, and Low sublevels. The levels of the ACTFL Guidelines (see Figure 2 below) describe a continuum of spoken proficiency from a highly articulate user capable of fully and spontaneously discussing topics at any level of abstraction down to a level of little or no functional ability. The ACTFL proficiency levels can be mapped onto the Common European Framework of Reference for Languages (CEFR) guidelines, a framework for assessing language skills commonly used in the European context (Assigning CEFR Ratings to ACTFL Assessments, 2016).

Table 7

Column 1: ACTFL proficiency sublevels; Column 2: excerpts from the ACTFL descriptors; Columns 3 and 4: comparison between the total number of participants rating on the pre- and posttests

ACTFL Level (Abbreviation)	ACTFL Descriptor Excerpt	Number of study participants at pretest	Number of study participants at posttest
Unratable	Responses that are completely in [a speaker's L1] or off-topic are considered unratable.	1	0
Novice Low (NL)	"no real functional ability… cannot therefore participate in a true conversational exchange"	32	17
Novice Mid (NM)	"communicate minimally by using a number of isolated words and memorized phrases… may say only two or three words at a time"	9	15
Novice High (NH)	"conversation is restricted to a few of the predictable topics necessary for survival… language consists primarily of short and sometimes incomplete sentences"	5	9
Intermediate Low (IL)	"able to handle successfully a limited number of uncomplicated communicative tasks… Conversation is restricted to some of the concrete exchanges and predictable topics necessary for survival"	3	6
Intermediate Mid (IM)	"able to handle successfully a variety of uncomplicated communicative tasks in straightforward social situations… related to self, family, home, daily activities… [etc.]"	3	4
Intermediate High (IH)	"able to handle successfully uncomplicated tasks and social situations requiring an exchange of basic information related to their work, school, … and areas of competence."	1	3

Note: All excerpts from ACTFL (2012, pp. 7-9).

The ACTFL Guidelines present the levels of proficiency as ranges, describing what an individual can and cannot do with language at each level, regardless of how, where and

when they learned. The following three case studies below contextualize the learning outcomes of some selected groups of participants who managed to improve their score over the course of the study in terms of the ACTFL Can-Do statements.

Table 9

ACTFL Can-Do Benchmarks. Oral proficiency in terms of Interpersonal Communication and Presentational Speaking at sublevels Novice Low – Intermediate Mid

ACTFL Level	Novice Low	Novice Mid	Novice High	Intermediate Low	Intermediate Mid
Interpersonal communication	I can communicate on some very familiar topics using single words and phrases that I have practiced and memorized.	I can communicate on very familiar topics using a variety of words and phrases that I have practiced and memorized.	I can communicate and exchange information about familiar topics using phrases and simple sentences, sometimes supported by memorized language. I can usually handle short social interactions in everyday situations by asking and answering simple questions.	I can participate in conversations on a number of familiar topics using simple sentences. I can handle short social interactions in everyday situations by asking and answering simple questions.	I can participate in conversations on familiar topics using sentences and series of sentences. I can handle short social interactions in everyday situations by asking and answering a variety of questions. I can usually say what I want to say about myself and my everyday life.
ACTFL Level	**Novice Low**	**Novice Mid**	**Novice High**	**Intermediate Low**	**Intermediate Mid**

Presentational speaking	I can present information about myself and some other very familiar topics using single words or memorized phrases.	I can present information about myself and some other very familiar topics using a variety of words, phrases, and memorized expressions.	I can present basic information on familiar topics using language I have practiced using phrases and simple sentences.	I can present information on most familiar topics using a series of simple sentences.	I can make presentations on a wide variety of familiar topics using connected sentences.

Case studies

The following case studies illustrate how three subsets of participants improved their oral proficiency and what communicative skills an L2 speaker has mastered at the respective level.

Case study 1: Learners who went from Novice Low to Novice Mid
One common learning outcome (N=12, or 22% of the study cohort) was an improvement from sublevel Novice Low to Novice Mid. The average study time to achieve this was 664 minutes, or approximately 11 hours or an average of 55 minutes per week over the course of the study. The least amount of Babbel study necessary to achieve this score gain was 4 hours, while the greatest amount of time required was just over 23 hours.

In terms of functional ability, Novice Mid represents a modest gain over Novice Low (see Figure 3 below). However, it does still represent some meaningful progress. L2 speakers assessed at the Novice Mid level of proficiency can, in principle, accomplish the following which Novice Low speakers cannot:

- I can introduce myself and provide basic personal information.
- I can answer a variety of simple questions.
- I can answer questions about what I am doing and what I did.
- I can answer questions about where I'm going or where I went.
- I can tell someone what I am doing.
- I can ask who, what, when and where questions
- I can communicate basic information about myself and people I know.
- I can give times, dates, and weather information.

Case study 2: Learners who went from Novice Low or Novice Mid to Novice High
Another frequent outcome on the OPIc tests (N=9 or 17% of the study cohort) was learners improving from sublevels Novice Low or Novice Mid to Novice High over the course of the study. These participants' average study time was 828 minutes; this equates to approximately 14 hours or an average of 69 minutes per week over the course of the study. The least amount of Babbel study necessary to achieve this score gain was 6.25 hours, while the greatest amount of time required was 25.7 hours.

In terms of functional ability, Novice High is a fair advancement over Novice Mid. L2 speakers assessed at the Novice High level of spoken proficiency can, in principle, accomplish the following tasks:

- I can exchange some personal information.
- I can ask and say someone's nationality.
- I can ask and talk about family members and their characteristics.
- I can ask for and give simple directions.
- I can make plans with others.
- I can invite and make plans with someone to do something or go somewhere.
- I can interact with others in everyday situations.
- I can order a meal and make a purchase.

Case study 3: Novice (all) to Intermediate Low or Mid

Unlike learners assessed at the novice level of proficiency, at the intermediate level of proficiency learners are able to "create with language when talking about familiar topics related to their daily life" (ACTFL, 2014). They can also "recombine learned material to express personal meaning." This represents a large leap in functional ability above the novice level. In this study, six learners (11% of the sample) was able to attain this with an average study time of 14 hours over the duration of the study. The least amount of Babbel study necessary to achieve this score gain was 6.7 hours, while the greatest amount of time required was just over 27 hours.

At the intermediate level of proficiency, learners would hypothetically be able to perform the following functions in Spanish:

- I can participate in conversations on familiar topics using sentences and series of sentences.
- I can handle short social interactions in everyday situations by asking and answering a variety of questions.
- I can ask for information, details, and explanations during a conversation.
- I can talk about my interests and hobbies.
- I can give some information about something I plan to do.
- I can talk about my favorite music, movies, and sports.
- I can use my language to handle tasks related to my personal needs.

Grammar Knowledge. The LMER model for grammar knowledge is summarized in Table 10. As with oral proficiency, previous classroom experience had a positive impact on initial level of grammar knowledge. A significant interaction of time and Babbel study hours was also found. No main effect or interaction involving Spanish interest was obtained. In practical terms, increases in Spanish grammar knowledge over time were dependent on Babbel study hours, with roughly two hours of study predicting a grammar score increase of 1 point.

Table 10.
Summary of grammar score linear mixed-effects regression model.

Fixed Effects	B (SE)	β	p

Intercept	4.90 (1.98)		
Time	2.81 (1.72)	0.10	0.11
Classroom Experience	2.96 (0.68)	0.44	<.001
Time x Babbel Study Hours	0.55 (0.13)	0.32	<.001

Random Effects	Variance (SD)
Participant (intercept)	90.13 (9.49)
Participant: Time (slope)	18.47 (4.30)

Model $R^2_{marginal}$ = .37, $R^2_{conditional}$ = .92. χ^2 vs. unconditional model = 81.98, df = 3, p <.001.

Figure 4 is a graphical representation of the LMER model for grammar knowledge. As with the oral proficiency model, the blue lines showing model predictions become progressively steeper as participants accumulate more Babbel study hours. There is also considerable variation in the intercepts and slopes of participants' lines. Unlike the oral proficiency plot, there are fewer flat lines, as the finer grain-size of the grammar test scores allow for smaller increases in grammar knowledge to be accounted for.

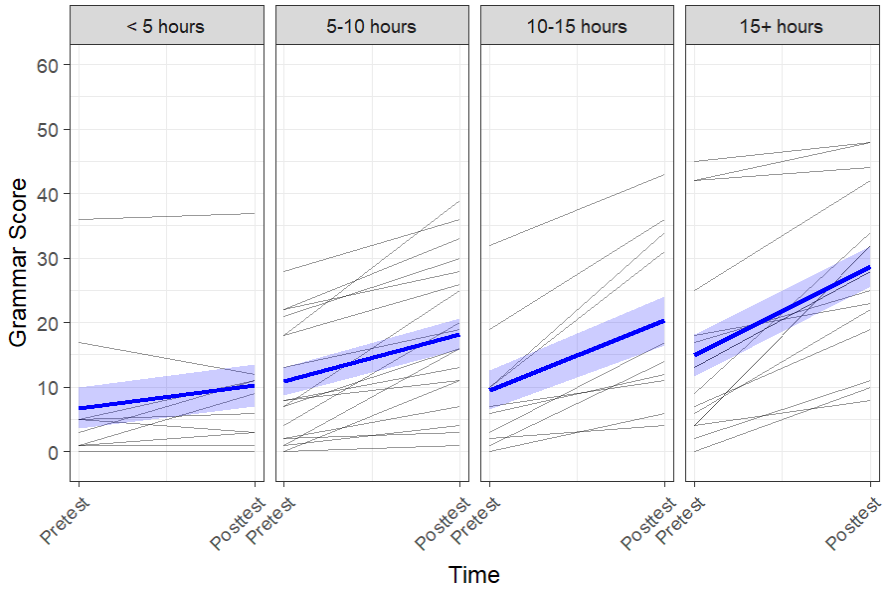


Figure 4. Changes in grammar knowledge over time according to number of hours studied on Babbel.

Vocabulary Knowledge. The LMER model for vocabulary knowledge is summarized in Table 8. Like oral proficiency and grammar knowledge, previous classroom experience had a positive impact on initial levels of grammar knowledge. A significant interaction of time and Babbel study hours was also found. No main effect or interaction involving Spanish interest was obtained. In practical terms, increases in Spanish vocabulary knowledge over time were dependent on total Babbel study hours, with roughly two hours of study predicting an increase of 1 point in vocabulary test score.

Table 11.
Summary of vocabulary score linear mixed-effects model.

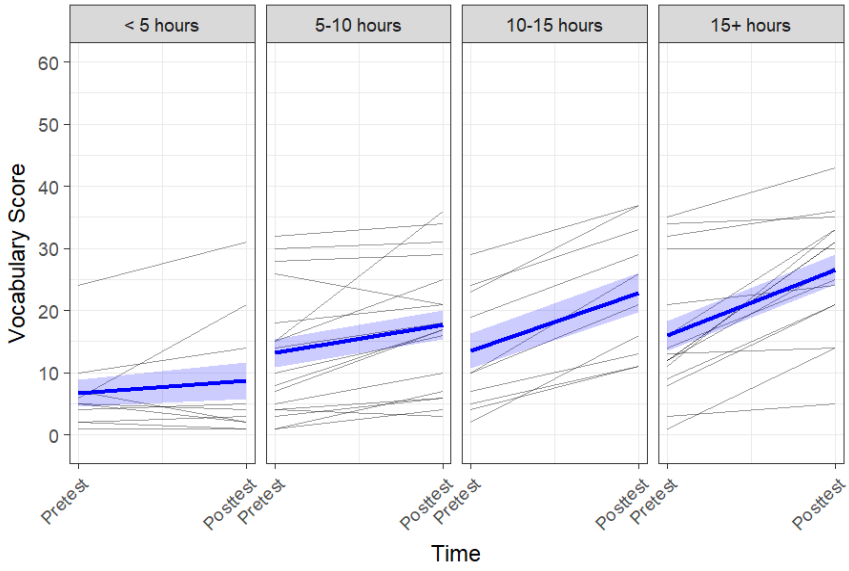
Fixed Effects	B (SE)	β	p
Intercept	6.58 (1.53)		
Time	1.17 (1.51)	0.05	0.44
Classroom Experience	3.04 (0.53)	0.54	<.001
Time x Babbel Study Hours	0.48 (0.11)	0.33	<.001

Random Effects	Variance (SD)
Participant (intercept)	55.21 (7.43)
Participant: Time (slope)	21.12 (4.60)

Model $R^2_{marginal}$ = .44, $R^2_{conditional}$ = .95. χ^2 vs. unconditional model = 80.234, df = 3, p <.001.

Figure 5 is a graphical representation of the LMER model for vocabulary knowledge. The blue lines showing model predictions in the right two panels are noticeably steeper than those in the left two panels. As seen in the plots for the previous two models, individual lines for vocabulary scores also demonstrated considerable variation in intercepts and slopes. There are very few lines with a negative slope; these may be attributable to some participants overconfidently indicating knowledge of the (quite plausible) nonwords when completing the posttest.

Figure 5. Changes in vocabulary knowledge over time according to number of hours studied on Babbel.



Discussion and Conclusions

After roughly three months of learning Spanish on Babbel, virtually all study participants made a gain in their language knowledge and/or ability to communicate.

Learning gains in terms of oral proficiency, grammar knowledge, and vocabulary knowledge were all associated with how much time participants spent on Babbel. The association between Babbel study hours and learning gains was stronger for grammar and vocabulary compared to oral proficiency. Given that Babbel presents limited opportunities for oral production as well as the broad-strokes measurement of oral proficiency provided by the OPIc, the fact that 59% of participants showed improvements and the connection between study hours and speaking gains are notable and encouraging. These findings show that Babbel's current pedagogical approach to mobile-assisted language learning allows learners to transfer receptive, input-based learning and explicit grammar and vocabulary instruction to communicative production (at Novice and Intermediate levels of oral proficiency, at least).

Some important considerations should be kept in mind when interpreting the results of this study. For one, the grammar and vocabulary measures we employed were deliberately linked to the Babbel Spanish curriculum, while the oral proficiency test was not. Thus, the stronger relationship between study hours and grammar/vocabulary gains compared to speaking gains is actually quite intuitive. Nonetheless, bringing in the ACTFL OPIc as a measure of oral proficiency, though inconsistent with the other curriculum-linked measures, has distinct benefits. The OPIc results are both more meaningful in terms of describing language ability (as opposed to a numerical score on a test used in a single research study) and directly comparable to external benchmarks and results from other studies utilizing the same measure. Explicit knowledge of discrete vocabulary items and grammar rules is easier to acquire--and to assess--than the type of implicit knowledge which supports communicative proficiency in an L2. Another important point to consider is that learner expectations for speaking development after short-term Babbel study for less than an hour a week (the average for this study) should be modest, yet positive. For example, while the benefits to speaking ability found in the present study are encouraging, for many participants (n = 12) the outcome was to go from "no real functional ability" (Novice Low) to being able to "communicate minimally by using a number of isolated words and memorized phrases" (Novice Mid). Finally, this study is limited in its generalizability due to the small sample size (cf. Vesselinov & Grego's 2016 study featuring 325 participants). A larger sample, particularly one featuring more learners with initial proficiency in ACTFL's Intermediate range, would lead to additional insights about the nature of learning Spanish on Babbel.

Despite the above mentioned limitations, this is a robust and methodologically rigorous study providing strong evidence that learning Spanish with Babbel facilitates the development of oral communication skills, and not only grammar and vocabulary acquisition. The study therefore makes an important contribution to the growing body of literature on commercial mobile-assisted language learning software and applications.

References

American Council on the Teaching of Foreign Languages. (2012). *ACTFL Proficiency Guidelines 2012.* Retrieved from

https://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf

American Council on the Teaching of Foreign Languages. (2016). *Assigning CEFR ratings to ACTFL assessments.* Retrieved from https://www.actfl.org/sites/default/files/reports/Assigning_CEFR_Ratings_To_ACTFL_Assessments.pdf

Cubbellotti, S. (2015). Examination evaluation of the ACTFL OPIc® in Arabic, English, and Spanish for the ACE Review. White Plains, NY: ACTFL. Retrieved from https://www.languagetesting.com/pub/media/wysiwyg/research/reports/Examination_Evaluation_of_the_ACTFL_OPIc_in_Arabic_English_and_Spanish_for_the_ACE_Review.pdf

Isbell, D. R., Winke, P., Gass, S. (in press). Using the ACTFL OPIc to assess proficiency and monitor progress in a tertiary foreign languages program. *Language Testing*.

Izura, C., Cuetos, F., & Brysbaert, M. (2014). Lextale-Esp: A test to rapidly and efficiently assess the Spanish vocabulary size. *Psicológica, 35,* 49-66.

Leonard, K. R., & Shea, C. E. (2017). L2 speaking development during study abroad: Fluency, accuracy, complexity, and underlying cognitive factors. *Modern Language Journal, 101*(1), 179-193. doi:10.1111/modl.12382

Lord, G. (2015). "I don't know how to use words in Spanish": Rosetta Stone and learner proficiency outcomes. *Modern Language Journal, 99*(2), 401-405.

Thompson, G. L., Cox, T. L., & Knapp, N. (2016). Comparing the OPI and OPIc: The effect of test method on oral proficiency scores and student preference. *Foreign Language Annals, 49*(1), 75-92. doi: 10.1111/flan.12178

Vesselinov, R., & Grego, J. (2016). The Babbel efficacy study [white paper]. Retrieved from https://press.babbel.com/en/releases/downloads/Babbel-Efficacy-Study.pdf