How well does Duolingo teach speaking skills?

Duolingo Research Report DRR-21-02
June 1, 2021 (12 pages)
https://www.duolingo.com/efficacy

Xiangying Jiang*, Joseph Rollinson*, Haoyu Chen*, Ben Reuveni*, Erin Gustafson*, Luke Plonsky†, and Bozena Pajak*

Abstract

Duolingo has previously been shown to be highly effective at teaching receptive listening and reading comprehension skills. The question remains as to how well Duolingo courses teach productive skills, such as speaking. This study measured the speaking proficiency of Duolingo learners who had completed the beginning-level course material in the Spanish and French courses. Results of the Pearson Versant Spanish Test and French Test showed that the speaking skills of Duolingo learners, who had little to no prior knowledge in the target language and used Duolingo as the only language learning tool, were in line with Duolingo's course expectations. Specifically, most of the study participants achieved the level of A2 or above on the CEFR scale. The findings of the study suggest that Duolingo is effective at teaching speaking in addition to listening and reading.

Keywords

Duolingo, efficacy, Spanish, French, speaking, foreign language

1 Introduction

Speaking has been deemed the most important but also the most difficult skill in language learning compared to reading, writing, and listening (Nunan, 2015). Most learners evaluate success in language learning by the ability to carry on a conversation in that language. To be able to speak, learners need to develop multiple sub-skills. These sub-skills include having sufficient vocabulary, knowing how to arrange words and phrases into sentences, using correct pronunciation, and more (Brown, 2018).

Duolingo is a language-teaching platform that offers free online courses available on mobile apps and the web. Duolingo has previously been shown to be highly effective at teaching receptive listening and reading comprehension skills (Jiang, Rollinson, Plonsky, & Pajak, 2020). The question remains as to how well Duolingo courses teach productive skills in a language. The current study addresses this gap by reporting speaking proficiency scores of Duolingo learners who had completed the beginning-level course material in the Spanish and French courses.

Over the years, academic researchers in language learning have expressed skepticism about the development of oral communicative abilities through app-based learning (Krashen, 2014; Lin & Warschauer, 2015; Loewen et al., 2019; Lord, 2015, 2016; Van Deusen-Scholl, 2015). For example, Lord (2015, 2016) found that beginning-level Spanish learners who used Rosetta Stone exclusively struggled in conversation compared to learners who received face-to-face instruction. Similarly, Loewen et al. (2019) found that beginning-level learners of Turkish on Duolingo did not do as well on oral tasks compared to tasks that targeted vocabulary and grammar. Loewen, Isbell, and Sporn (2020) reported similar results with 54 learners who spent an average of 12 hours learning Spanish on Babbel during a period of three months. The study showed that all learners gained in vocabulary and grammar but only 59% of them improved in speaking as assessed by the computer version of the Oral Proficiency Interview (OPIc) offered by the American Council on the Teaching of Foreign Languages (ACTFL). The ACTFL proficiency scale has ten sublevels, ranging from Novice (low, mid, high) to Intermediate (low, mid, high), Advanced (low, mid, high), and Superior. Like in many other studies (e.g., Isbell, Winke, & Gass, 2019; Rubio & Hacking, 2019; Tschirner, 2016), the researchers converted ACTFL sublevels into integers for quantitative analysis, mapping Novice Low to 1, Novice Mid to 2, and so on. On average, the ACTFL sublevels of the participants went from 1.81 (approaching Novice Mid) in pretest to 2.52 (Novice Mid) in posttest. The researchers indicated that the speaking gains were "modest" (p. 19). According to ACTFL, at the level of Novice Mid, learners are able to "communicate minimally by using a number of isolated words and memorized phrases" (ACTFL, 2012). Loewen et al. (2020) acknowledged that "any gains (on speaking ability) are encouraging" but suggested "tempered interpretations of the magnitude of oral

*Duolingo, Inc.
†Northern Arizona University

Corresponding author:
Xiangying Jiang
Duolingo, Inc. 5900 Penn Ave
Pittsburgh, PA 15206, USA
Email: assessment-study@duolingo.com

proficiency growth exhibited by most learners in this study" (p. 19).

The findings of Vesselinov and Grego (2016) also aligned with Loewen et al. (2020). Vesselinov and Grego (2016) assessed 61 learners who spent an average of 24 hours learning Spanish on Busuu during a period of two months. The study found that 75% of the learners showed improvements in oral proficiency as assessed by ACTFL OPI. In order to compare results across studies, we converted the pretest and posttest ACTFL ratings in Vesselinov and Grego (2016) to integers in the same way as Loewen et al. (2020). This analysis showed that on average the Busuu learners in Vesselinov and Grego (2016) improved from 1.49 (Novice Low) at pretest to 2.66 (Novice Mid) at posttest. This improvement is greater than the progress reported in Loewen et al. (2020) and may be due to learners having spent double the number of hours of Babbel learners (24 hours vs. 12 hours) in a shorter period of time (2 months vs. 3 months).

The goal of the current study was to measure the speaking proficiency of Duolingo learners who had completed the beginning-level course material in the Spanish and French courses. In particular, the current study aimed to answer the following research questions:

1. What levels of speaking proficiency do Duolingo learners achieve upon completing the beginning-level units of the Spanish or French course?
2. To what extent do Duolingo Spanish and French courses develop learners' abilities in the sub-skills of speaking, including sentence mastery, vocabulary, fluency, and pronunciation?

Before explaining the methods of data collection, we provide a brief description of the Duolingo course structure and the standards that guide course development at Duolingo.

1.1 The Beginning-level Units of the Spanish and French Courses

The beginning-level content of a Duolingo course includes five units, each of which concludes with a "checkpoint" (see Figure 1). Each circle in Figure 1 represents a skill, which is a collection of lessons on either a communicatively functional topic (such as travel-related vocabulary and expressions, or ordering at a restaurant) or a grammar-focused topic (such as present tense conjugation or pronouns). There are a total of 114 skills on functional topics and 15 grammar skills in the beginning-level units of the Spanish course (from the English user interface). The beginning-level units of the French course (from the English user interface) includes 99 skills on functional topics and 19 grammar skills (see Table 1). Each skill on a functional topic includes 5 difficulty levels and each grammar skill has 2 levels. There are 4-5 lessons at each level. Learners are required to complete at least one difficulty level in each skill in a row to unlock the next row, but they can choose whether to complete more levels or not. As a result, there can be substantial variation among individual learners on the percentage of content they complete before reaching the end of Unit 5.

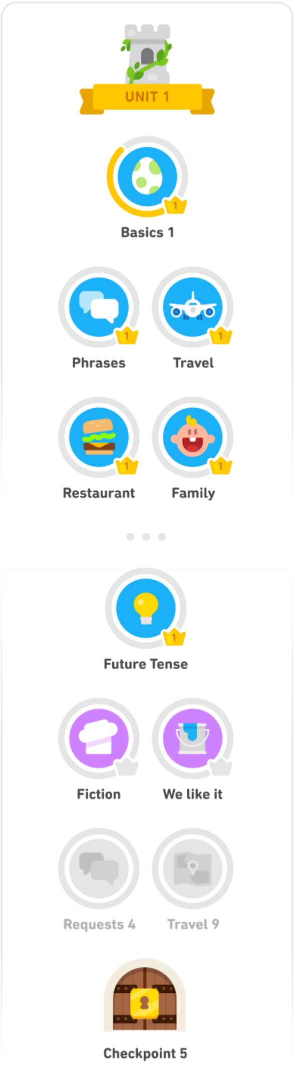

Figure 1. Example Duolingo course structure.

The Duolingo course units are aligned with the Common European Framework of Reference (CEFR), an international standard for describing the abilities of language learners at various stages of proficiency. See Table 1 for the CEFR level of each course unit. The CEFR divides language proficiency into three broad levels–A (Basic User), B (Independent User), and C (Proficient User), which correspond to the traditional beginner, intermediate, and advanced levels (Council of Europe, 2001). Each broad level is then further divided into two levels, namely, A1 and A2, B1 and B2, and C1 and C2 (see Figure 2).

Table 1. Number of Units in Each Section of the Duolingo Spanish and French Courses

Course unit	CEFR alignment	Spanish: # of skills	French: # of skills
1	Pre-A1	9	12
2	A1.1	29	29
3	A1.2	32	25
4	A2.1	29	28
5	A2.2	30	24
Total		129 (15 grammar skills)	118 (19 grammar skills)

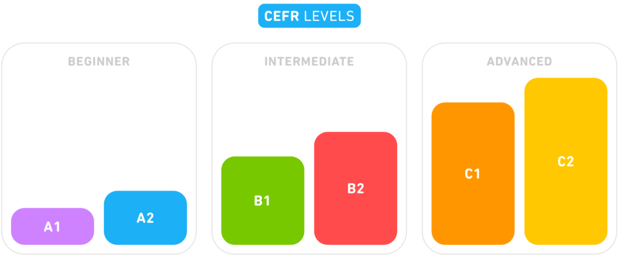

Figure 2. CEFR levels.

The Duolingo beginning-level course sections, units 1-5, correspond to A1-A2 in the CEFR, which means that the expected proficiency of learners who complete Unit 5 is at the A2 level. In the area of spoken language use, CEFR (Council of Europe, 2001, p. 29) provides guidance about A2 speaking skills across five different aspects of speaking: range, accuracy, fluency, interaction, and coherence. See Table 2.

Duolingo lessons include several activity types targeting learning and practice in vocabulary, grammar, reading, listening, and speaking. To facilitate listening and speaking development, Duolingo provides learners with many opportunities to listen to the target language and speak it out loud. All Duolingo Spanish and French course content is accompanied by audio and learners are allowed to play the audio at varied speeds as often as they need. In addition, speech recognition technology is used for all speaking exercises in both courses in order to provide learners with feedback.

In the next section, we explain in detail how the study was conducted.

2 Methods

2.1 Participants

The participants of the study were 156 Spanish and 102 French learners on Duolingo who studied these languages from the English user interface. These learners had to meet the following selection criteria to be included in the study. The participants were:

1. learners whose IP addresses were not in countries or regions where Spanish or French is an official or widely spoken language (see Appendix A for the list of countries and regions). By doing so, we excluded learners who studied the target language while immersed in the target-language culture; this was important because the development of speaking skills is sensitive to exposure to the target language in the environment (Alptekin, 1983; Klein & Dimroth, 2009; Perdue, 2002).
2. learners who self-reported using Duolingo as the only language learning tool. They confirmed that they did not take classes or use other programs or apps during their Duolingo course.
3. learners who had self-reported having no or little prior proficiency in the target language prior to beginning the Duolingo course. In particular, we included only those learners who reported prior proficiency of 0-2 on a 0-10 scale, with 0 representing "I have no knowledge of the language at all," and 10 indicating "I have perfect knowledge of the language." Note that Duolingo collects this information from all learners upon completion of Unit 1 for the purposes of learner analytics and not for course placement.
4. learners who reached the end of Unit 5 within the data collection window. This point marks the completion of the beginning-level course content on Duolingo.
5. learners aged 18 or older.

Participants completed a background survey about their language history, education, and motivation; for more details about the survey, see the Instruments section below. In Table 3 we summarize the demographic characteristics of the participants. Overall, the participants in the two courses had similar characteristics, except that the participants in French were slightly younger and more likely to study the language for job-related purposes, while the participants in Spanish were somewhat older and more likely to study the language for travel and social purposes.

3 Instruments

Table 2. Descriptions of CEFR A2 Speaking Proficiency

Aspects of speaking	Can-do statements
Range	Can use basic sentence patterns with memorized phrases, groups of a few words and formulae in order to communicate limited information in simple everyday situations.
Accuracy	Can use some simple structures correctly, but still systematically makes basic mistakes.
Fluency	Can make him/herself understood in very short utterances, even though pauses, false starts, and reformulation are very evident.
Interaction	Can answer questions and respond to simple statements; can indicate when he/she is following but is rarely able to understand enough to keep conversation going of his/her own accord.
Coherence	Can link groups of words with simple connectors like "and," "but," and "because."

Table 3. Characteristics of the Participants

Characteristics	Spanish	French
Age		
18-34 years	48.72%	63.73%
35-54 years	30.77%	22.55%
55-74 years	20.51%	13.73%
Home language before age 6		
Only English	47.40%	44.12%
Only one language, but not English or the target language assessed in the study	42.86%	47.06%
More than one language, but not the target language assessed in the study	9.74%	8.82%
Highest level of education		
Bachelor's degree	46.15%	45.10%
Master's degree	33.33%	27.45%
Doctoral degree	8.33%	11.76%
Other	12.18%	15.69%
Primary reason for learning the language		
For fun/leisure	48.08%	53.92%
For travel	37.18%	30.39%
For memory / brain acuteness	18.59%	22.55%
For job-related purposes	14.10%	24.51%
For social purposes	21.15%	14.71%
For school	5.77%	9.80%
Other	15.38%	12.75%

3.1 The Background Survey

The background questionnaire included questions related to participants' language background, reasons for learning the language, level of education, age group, and whether they took classes or used other programs/apps during the time they used Duolingo. The latter question confirmed eligibility to satisfy Criterion #2 for participant selection; see Participants above.

3.2 The Versant Spanish and French Tests

The Versant Spanish and French Tests are tests of spoken language developed by Pearson Education (https://www.pearson.com/english/versant.html). The spoken language tests were designed to "measure the core speaking skills" of language learners. The Spanish and French tests include seven tasks. Table 4 lists each task with a brief description (Pearson Education, 2019).

The Versant Spanish and French Tests require test-takers to read sentences aloud, listen and repeat sentences, say the opposites of words they hear, answer short questions, build sentences from jumbled-up word combinations, retell stories, and answer open-ended questions. According to the test description (Pearson Education, 2018a, 2018b), the tests place a great deal of emphasis on automaticity with the language. In particular, the demands for automaticity are shown in tasks such as "saying the opposite of a word you hear" and "building sentences from jumbled-up word combinations you hear." For these tasks, test-takers are required to recognize words or word combinations they hear, quickly access and retrieve lexical items or build

Table 4. Tasks in the Versant Spanish and French Tests

Part	Task	Description
A	Read Sentences	See a sentence on the test screen and read it aloud.
B	Repeat sentences	Hear a sentence and repeat it.
C	Say the opposites	Hear a word and say its opposite.
D	Answer questions	Give a simple answer to a question.
E	Build sentences	Hear jumbled-up word groups, rearrange them into a sentence, and speak it.
F	Retell stories	Hear a brief story and retell it.
G	Respond to open-ended questions	Hear a question prompt and answer it within 30 seconds.

phrases and clause structures, and articulate them under extreme time pressure.

The responses from test-takers are scored automatically by means of a speech recognition and parser program. The score report (Pearson Education, 2019–2020) provides an overall proficiency score and four subscores (fluency, pronunciation, sentence mastery, vocabulary), all scored between 20-80. The overall score of the test represents the ability to understand the spoken language and "speak it intelligibly at a native-like conversational pace on everyday topics" (Pearson Education, 2018b, p. 11), and it is calculated based on a weighted combination of the four diagnostic subscores (30% Sentence Mastery, 20% Vocabulary, 30% Fluency, and 20% Pronunciation). Among the four subcomponents of speaking, sentence mastery measures "the ability to understand, recall, and produce phrases and clauses in complete sentences"; vocabulary "reflects the ability to understand common everyday words spoken in sentence context and to produce such words as needed"; fluency is measured from "the rhythm, phrasing and timing evident in constructing, reading and repeating sentences"; and pronunciation assesses "the ability to produce consonants, vowels, and stress in a native-like manner in sentence context" (Pearson Education, 2018b, pp. 11–12).

Based on the Test Description and Validation Summary (Pearson Education, 2018a, 2018b), the split-half reliability coefficients of the Spanish test and the French test were both 0.97, indicating that both tests are highly reliable. The split-half reliability coefficients for the Spanish subscores ranged from 0.91 to 0.95 and those for the French subscores were 0.77 for vocabulary, 0.89 for sentence mastery, 0.93 for fluency, and 0.95 for pronunciation. Furthermore, the Versant Spanish and French scores correlate with CEFR estimates at 0.90 and 0.88, respectively. The overall score and the subscores are mapped to the CEFR scales as shown in Table 5, with detailed oral interaction descriptors in Appendix B (Pearson Education, 2018a, 2018b).

The Versant test takes 15-17 minutes to complete. To strengthen the validity of our findings, we used the remote monitoring feature provided by HirePro. It video-records participants as they take the test to flag suspicious behavior (e.g, a second person entering the camera view) and monitor browser use (to see if they navigate away from the test). In the data we report in our analysis, we excluded all scores from participants who were marked "suspicious" by the system (see Table 6 for the number of suspicious scores).

Table 5. Mapping of Versant Spanish and French Test Scores to CEFR Levels

Versant test score	CEFR level
79-80	C2
69-78	C1
58-68	B2
47-57	B1
36-46	A2
26-35	A1
20-25	<A1

4 Procedures

We sent an email soliciting participation in the research study to a random sample of Duolingo learners when they completed Unit 5 in the Spanish or French course, if they met the following criteria: prior proficiency of 0-2 in the language and an IP address in countries where Spanish or French is not an official or widely spoken language. Learners aged 18 and above who were interested in participating completed a background survey to verify eligibility and collect additional demographic information. Learners who responded that they had taken classes or used other programs/apps to learn the language during the time they used Duolingo were excluded from participation.

Qualified participants were emailed on a rolling basis and invited to take the Versant Spanish or French Test for free. Participants completed the test within two weeks. Each participant received $20 and their score report after taking the test. Table 6 shows the data collection funnel. This funnel is noteworthy in several respects. First, only about 40% of the learners who were eligible for the test attempted to take the test. This large drop in participation rate was mostly due to lack of appropriate equipment. The incorporation of the remote monitoring system imposed restrictions and high system requirements; for example, it only allowed the test to be taken with Version 80 or higher of the Google Chrome browser on a computer with a stable internet connection and high quality video and audio equipment. Most Duolingo learners use their mobile phones to learn and communicate with Duolingo and they might not have access to all the required equipment. Second, 18 learners in French and 20 in Spanish started the test but did not complete it. Third, 43 participants (about 28%) in French did not receive a score after they completed the test. According to a Versant representative, "this seemed to suggest that some candidates are either not speaking clearly in French or are taking the test in an improper environment (background noise noise, faulty mic, etc.)" (M. Kumar, personal communication, May 18, 2021). However, given that the same was not the case with our participants in Spanish, the improper environment explanation seems less likely; instead, the pronunciation of the participants in French was probably insufficiently clear for the Versant French speech recognition program. Finally, the remote monitoring system detected suspicious behaviors of 10 participants in French and 19 participants in Spanish. These suspicious scores were excluded in the following analyses.

5 Results

To answer the first research question–what levels of speaking proficiency did Duolingo learners achieve upon completing the beginning-level course content for Spanish or French–we report the means and standard deviations of the overall scores on the Versant test (see Table 7). According to the guidelines for mapping Versant scores to CEFR levels (see Table 5), a score range of 36-46 indicates the CEFR level of A2. Therefore, for Duolingo Spanish learners, an average of 40.97 indicates solid A2 speaking abilities. For Duolingo French learners, an average of 36.72 indicates a low A2.

In addition to learners' average scores, we also present the distribution of scores in Figure 3. For Spanish, 66.03% of learners scored at A2 or above; for French, 52.94% of learners scored at A2 or above.

To answer our second research question concerning the extent to which the Duolingo Spanish and French courses prepare learners in the sub-skills of speaking, including sentence mastery, vocabulary, fluency, and pronunciation, we report the means and standard deviations of the subscores in Table 8 and then Figure 4.

The subscores provide important diagnostic feedback regarding Duolingo courses. First, there were dramatic differences in pronunciation scores across the Spanish and French learners: the Spanish pronunciation score was the highest of all subscores, while the French pronunciation score was the weakest of all subscores in both courses and lowered the overall French score. The French pronunciation score (30.37) fell below the A2 threshold of 36 and indicates that more and improved pronunciation instruction is needed to meet the goal of teaching A2-level pronunciation skills by the end of Unit 5 in the Duolingo French course. The fact that 43 (28%) completed tests were unable to be scored by the speech recognizer (see Table 6) could be additional evidence that the French participants struggled with pronunciation. Although the Versant representative could not confirm the exact reason why these tests were not scored, it is unlikely this was due to recording quality since there were no similar problems with Spanish tests. If the unscored tests were indeed due to low participant intelligibility caused by poor French pronunciation, including all these participants in our sample could have further lowered the already low French pronunciation scores.

On average, participants in both Spanish and French Duolingo courses demonstrated A2 speaking abilities in the sub-skills of sentence mastery and fluency. Duolingo courses focus on sentence-level language throughout all lessons and levels, so the participants had a considerable amount of practice building sentences in their Duolingo exercises. Participants in both courses achieved, on average, solid A2 scores in understanding, recalling, and producing phrases and clauses in complete sentences, as measured by the sentence mastery subscore. They also achieved, on average, A2 level for the fluency subscore, which measured their ability in producing rhythmic language and appropriate phrasing in constructing, reading, and repeating sentences.

The subscores also showed that the vocabulary score was the weakest in Spanish and the second weakest in French. The lower vocabulary scores might have been related to the specific test tasks in the Versant Spanish and French Tests. For example, one of the tasks that assesses vocabulary knowledge asks the test-takers to say the opposites of the words they hear within a few seconds. This task requires strong automaticity in lexical access and retrieval (Pearson Education, 2018a, 2018b), which exerts high time pressure on the test-takers. A less time-sensitive measure of vocabulary knowledge (i.e., one that relies less on automatic production) would have likely yielded higher scores in this domain. Duolingo courses, however, may be more facilitative in developing learners' receptive vocabulary knowledge. Having more activities in the courses that require lexical retrieval in productive tasks would likely be beneficial for Duolingo learners, especially in timed vocabulary tasks such as the one used in the Versant Spanish and French Tests.

6 Discussion

This study evaluated the speaking proficiency of Duolingo learners who had completed the beginning-level course content in the Spanish and French courses. The results of the study showed that, on average, the participants in the Spanish course achieved solid A2 speaking abilities and those in French achieved a somewhat weaker A2 level. Specifically, about two-thirds of the participants (66%) in Spanish and more than half

Table 6. Data Collection Funnel

	Email sent	Survey responded	Test eligible	Test started	Test completed	Test successfully scored	Test valid (non-suspicious)
Spanish	8367	813	499	195	175	175	156
French	3177	815	478	173	155	112	102

Table 7. Means and Standard Deviations of Overall Versant Scores of Duolingo Learners

Course	N	Mean	Standard deviation
Spanish	156	40.97	11.95
French	102	36.72	8.48

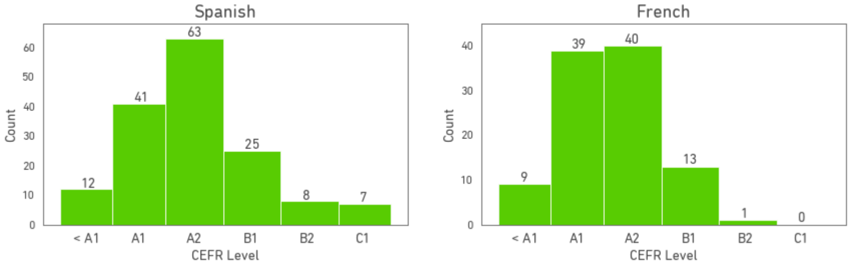

Figure 3. Distribution of Versant test scores of Duolingo learners based on CEFR levels.

Table 8. Means (and Standard Deviations) of Versant Test Subscores of Duolingo Learners

Course	N	Sentence mastery	Vocabulary	Fluency	Pronunciation
Spanish	156	42.51 (16.70)	37.33 (13.04)	39.52 (12.27)	45.67 (9.46)
French	102	39.39 (10.24)	36.14 (9.25)	38.96 (10.75)	30.37 (8.31)

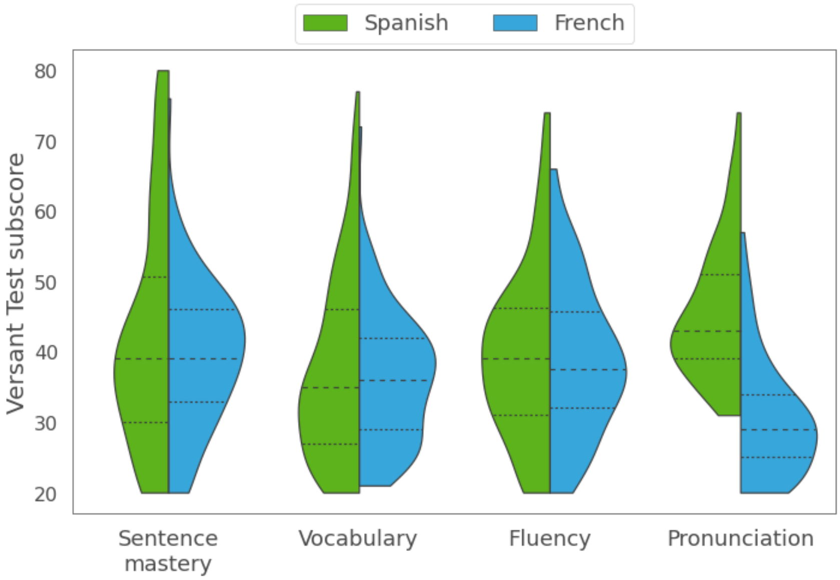

Figure 4. Distribution of Versant Test subscores of Duolingo learners shown with density plots. Dashed line represents median and dotted lines represent interquartile range.

of the participants (53%) in French achieved the level of A2 or above in speaking. With extensive opportunities in the Duolingo courses to hear the target language and practice sentence-level speaking with feedback from the speech recognition program, participants developed speaking skills and reached the proficiency level targeted by the CEFR-based curriculum standards.

The subscores of the speaking tests were mostly in line with the overall scores, but three observations are noteworthy. First, the subscores of the speaking tests indicated a strong contrast on pronunciation skills of the participants in Spanish and French. Among the subscores, pronunciation scored the highest in Spanish but the lowest in French. This is not entirely surprising given that French pronunciation is known to be difficult for English speakers to learn, requiring pedagogical attention over a long time (Huensch, 2019; Sturm, 2019). Second, the subcores on sentence mastery and fluency indicated that Duolingo learners developed these sub-skills as expected, and they were able to understand, recall, and produce complete sentences, and articulate them with good rhythm and appropriate phrasing. The third observation is that the vocabulary subscores were relatively low for both Spanish and French learners. Previous research has consistently shown that vocabulary learning is one of the strengths of mobile-based language learning (Loewen et al., 2020; Lord, 2015, 2016; Vesselinov & Grego, 2016). However, this study demonstrates that Duolingo's emphasis on receptive vocabulary knowledge may not transfer directly to productive knowledge, especially when automaticity is the goal of the assessment. At the same time, these results are not necessarily an indication of a lack of vocabulary knowledge among beginner Duolingo users. Rather, the relatively low vocabulary subscore may be seen in part as an artifact of Versant's vocabulary measure which requires a high level of automaticity in speech production.

The results for learners in the French course, however, need to be taken with caution. As mentioned earlier, about 28% of the test-takers in French did not receive a score after they completed the test, but that was not the case for Spanish participants. Although we cannot pinpoint the exact reason, it is likely that the French participants did not speak clearly enough for the scoring system to capture meaningful production in French. If that was the case, their inclusion would have lowered the overall average score for French learners. We consider this an important limitation of our findings.

The speaking assessment provided important diagnostic information to help us understand the strengths and weaknesses of Duolingo courses in teaching various components of learners' speaking ability. One pedagogical implication of the findings is the need to enhance the intelligibility of Duolingo learners by teaching pronunciation more effectively in French (Hirschi, 2020). Another pedagogical implication of the findings is that in addition to teaching receptive vocabulary knowledge, Duolingo courses would benefit from more activities that would facilitate the development of productive vocabulary knowledge.

7 Conclusion

The results of the speaking assessment demonstrated that most beginning-level Duolingo learners have achieved the expected proficiency outcomes and curriculum objectives for speaking skills. Specifically, the test subscores indicated that Duolingo learners have speaking abilities in line with the standards for four speaking sub-skills, with the exception of French learners' pronunciation. Together with findings from a previous study (Jiang et al., 2020) which evaluated the listening and reading proficiency of Duolingo learners, this study complements the accumulating body of evidence of the efficacy of the beginning-level course content in the Duolingo Spanish and French courses.

Author Biographies

Xiangying Jiang is a lead learning scientist and works on learning assessment at Duolingo. She has a PhD in Applied Linguistics and was Associate Professor of TESOL at West Virginia University before joining Duolingo.

Joseph Rollinson is currently a staff software engineer at Duolingo, where he co-leads teams focused on learning assessment and learning infrastructure. He graduated from Carnegie Mellon University with undergraduate degrees in Computer Science and Philosophy. As an undergraduate, he performed research in intelligent tutoring systems.

Haoyu Chen is currently a software engineer at Duolingo. She graduated from University of Pennsylvania with a Master's degree in Data Science.

Ben Reuveni holds a Ph.D in Cognitive Neuroscience (Northwestern University, 2020). His research focused primarily on computational cognitive models of explicit-implicit interactions during decision making. He is currently a learning scientist and works on learning assessment and feature development.

Erin Gustafson holds a PhD in Linguistics (Northwestern University, 2016). Prior to joining Duolingo in 2017, Erin was a post-doctoral fellow at the Northwestern University Medical School, where her research focused on machine learning and natural language processing applications in the medical domain. Her graduate research focused on bilingualism and psycholinguistics. She is currently Lead Data Scientist at Duolingo and co-leads a team focused on learning assessment.

Luke Plonsky is Associate Professor of Applied Linguistics at Northern Arizona University. His work, focusing primarily on second-language acquisition and research methods, has appeared in over 80 articles, book chapters, and books. Luke is Associate Editor of *Studies in Second Language Acquisition*, Managing Editor of *Foreign Language Annals*, and Co-Director of the IRIS Database.

Bozena Pajak holds a Ph.D. in Linguistics (University of California, San Diego, 2012). Before joining Duolingo in 2015, she was a Research Associate and a Lecturer in Linguistics at Northwestern University. Her research focused primarily on the acquisition of additional languages in adulthood. She is currently the Director of Learning and Curriculum at Duolingo, where she co-leads the company's Learning Area.

8 References

ACTFL. (2012). *ACTFL proficiency guidelines 2012*. Retrieved from https://www.actfl.org/sites/default/files/guidelines/ACTFLProficiencyGuidelines2012.pdf

Alptekin, C. (1983). Target language acquisition through acculturation: EFL learners in the english-speaking environment. *The Canadian Modern Language Review*, *39*, 818–826.

Brown, H. D. (2018). *Language assessment: Principles and classroom practices. (3rd. ed)*. White Plains, NY: Pearson Education.

Council of Europe. (2001). *Common european framework of references for languages: Learning, teaching, assessment*. Retrieved from https://rm.coe.int/1680459f97

Hirschi, K. (2020). Duolingo [review]. In O. Kang, S. Staples, K. Yaw, & K. Hirschi (Eds.), *Proceedings of the 11th pronunciation in second language learning and teaching conference* (pp. 354–359). Ames, IA: Iowa State University.

Huensch, A. (2019). The pronunciation teaching practices of university-level graduate teaching assistants of french and spanish introductory language courses. *Foreign Language Annals*, *52*, 13–31. https://doi.org/10.1111/flan.12372

Isbell, D. R., Winke, P. M., & Gass, S. M. (2019). Using the ACTFL OPIc to assess proficiency and monitor progress in a tertiary foreign languages program. *Language Testing*, *36*, 439–465. https://doi.org/10.1177/0265532218798139

Jiang, X., Rollinson, J., Plonsky, L., & Pajak, B. (2020). *Duolingo efficacy study: Beginning-level courses equivalent to four university semesters [white paper]*. Retrieved from https://duolingo-papers.s3.amazonaws.com/reports/duolingo-efficacy-whitepaper.pdf

Klein, W., & Dimroth, C. (2009). Untutored second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *The new handbook of second language acquisition* (pp. 503–522). Bingley, UK: Emerald Group.

Krashen, S. (2014). Does duolingo "trump" university-level language learning? *The International Journal of Foreign Language Teaching*, 13–15. Retrieved from http://sdkrashen.com/content/articles/krashen-does-duolingo-trump.pdf

Lin, C.-H., & Warschauer, M. (2015). Online foreign language education: What are the proficiency outcomes? *The Modern Language Journal*, *99*, 394–397. https://doi.org/10.1111/modl.12234_1

Loewen, S., Crowther, D., Isbell, D. R., Kim, K. M., Maloney, J., Miller, Z. F., & Rawal, H. (2019). Mobile-assisted language learning: A duolingo case study. *ReCALL*, *31*, 293–311. https://doi.org/10.1017/S0958344019000065

Loewen, S., Isbell, D. R., & Sporn, Z. (2020). The effectiveness of app-based language instruction for developing receptive linguistic knowledge and oral communicative ability. *Foreign Language Annals*, *53*, 209–233.

Lord, G. (2015). "I don't know how to use words in spanish": Rosetta stone and learner proficiency outcomes. *The Modern Language Journal*, *99*, 401–405.

Lord, G. (2016). Rosetta stone for language learning. *IALLT Journal of Language Learning Technologies*, *46*, 1–35.

Nunan, D. (2015). *Teaching english to speakers of other languages: An introduction*. New York, NY: Routledge.

Pearson Education. (2018a). *Versant French Test: Test description and validation summary*. Retrieved from https://www.pearson.com/content/dam/one-dot-com/one-dot-com/english/SupportingDocs/Versant/ValidationSummary/Versant-French-Test-Description-Validation-Summary.pdf

Pearson Education. (2018b). *Versant Spanish Test: Test description and validation summary*. Retrieved from https://www.pearson.com/content/dam/one-dot-com/one-dot-com/english/SupportingDocs/Versant/ValidationSummary/Versant-Spanish-Test-Description-Validation-Summary.pdf%20%0A

Pearson Education. (2019). *Versant Spanish Test: Sample test paper*.

Pearson Education. (2019–2020). *Versant Spanish Test: Sample score report*. Retrieved from https://www.pearson.com/content/dam/one-dot-com/one-dot-com/english/versant-test/Versant_Spanish_Test.pdf

Perdue, C. (2002). Development of L2 functional use. In V. Cook (Ed.), *Portraits of the L2 user* (pp. 121–144). Clevedon, UK: Multilingual Matters.

Rubio, F., & Hacking, J. F. (2019). Proficiency vs. Performance: What do the tests show? In P. Winke & S. Gass (Eds.), *Foreign language proficiency in higher education* (pp. 137–152). Springer International Publishing AG. Retrieved from http://ebookcentral.proquest.com/lib/wvu/detail.action?docID=5622543

Sturm, J. L. (2019). Current approaches to pronunciation instruction: A longitudinal case in french. *Foreign Language Annals*, *52*, 32–44. https://doi.org/10.1111/flan.12376

Tschirner, E. (2016). Listening and reading proficiency levels of college students. *Foreign Language Annals*, *49*, 201–223. https://doi.org/10.1111/flan.12198

Van Deusen-Scholl, N. (2015). Assessing outcomes in online foreign language education: What are key measures for success? *The Modern Language Journal*, *99*, 398–400.

Vesselinov, R., & Grego, J. (2016). *The Busuu efficacy study [white paper]*. Retrieved from https://blog.busuu.com/wp-content/uploads/2016/05/The_busuu_Study2016.pdf

A Appendix

Table 9. Spanish- and French-Speaking Countries or Regions. Duolingo learners whose IP addresses were in those countries or regions (Spanish-speaking for learners of Spanish, and French-speaking for learners of French) were considered ineligible to participate in this study.

Spanish-speaking countries or regions	Argentina; Belize; Bolivia; Chile; Colombia; Costa-Rica; Cuba; Dominican-Rep; Ecuador; El-Salvador; Equatorial-Guinea; Gibraltar; Guatemala; Honduras; Mexico; Morocco; Nicaragua; Panama; Paraguay; Peru; Puerto-Rico; Spain; Uruguay; Venezuela; Western-Sahara
French-speaking countries or regions	Algeria; Belgium; Benin; Burkina-Faso; Burundi; Cambodia; Cameroon; Canada; Central-African-Rep; Chad; Comoros; Congo; Cote-dIvoire; Democratic-Rep-Congo; Djibouti; Dominica; Equatorial-Guinea; France; French-Guiana; French-Polynesia; Gabon; Guadeloupe; Guinea; Haiti; Jersey; Lao; Lebanon; Luxembourg; Madagascar; Mali; Martinique; Mauritius; Mayotte; Monaco; Morocco; Niger; New-Caledonia; Reunion; Rwanda; Senegal; Seychelles; St-Barthelemy; St-Lucia; St-Martin-Fr; St-Pierre-Miquelon; Switzerland; Togo; Tunisia; Vanuatu; Vatican; Wallis-Futuna; Western-Sahara

B Appendix

Table 10. Relation of Scores of Versant Spanish and French Tests to Oral Interaction Descriptors Based on Council of Europe (2001) Framework (as cited in Pearson Education, 2018a, 2018b)

Versant Spanish or French Test Score	CEFR level	Oral Interaction Descriptors Based on Council of Europe (2001)
79-80	C2	Conveys finer shades of meaning precisely and naturally. Can express him/herself spontaneously at length with a natural colloquial flow. Consistent grammatical and phonological control of a wide range of complex language, including appropriate use of connectors and other cohesive devices.
69-78	C1	Shows fluent, spontaneous expression in clear, well-structured speech. Can express him/herself fluently and spontaneously, almost effortlessly, with a smooth flow of language. Clear, natural pronunciation. Can vary intonation and stress for emphasis. High degree of accuracy; errors are rare. Controlled use of connectors and cohesive devices.
58-68	B2	Relates information and points of view clearly and without noticeable strain. Can produce stretches of language with a fairly even tempo; few noticeably long pauses. Clear pronunciation and intonation. Does not make errors that cause misunderstanding. Clear, coherent, linked discourse, though there may be some "jumpiness."
47-57	B1	Relates comprehensibly main points he/she wants to make on familiar matters. Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair may be very evident. Pronunciation is intelligible even if a foreign accent is sometimes evident and occasional mispronunciations occur. Reasonably accurate use of main repertoire associated with more predictable situations. Can link discrete, simple elements into a connected sequence.
36-46	A2	Relates basic information on, e.g., work, background, family, free time, etc. Can make him/herself understood in very short utterances, even though pauses, false starts, and reformulation are very evident. Pronunciation is generally clear enough to be understood despite a noticeable foreign accent. Uses some simple structures correctly, but still systematically makes basic mistakes. Can link groups of words with simple connectors like "and," "but," and "because."
26-35	A1	Makes simple statements on personal details and very familiar topics. Can manage very short, isolated, mainly prepackaged utterances. Much pausing to search for expressions to articulate less familiar words. Pronunciation is very foreign.
20-25	<A1	Candidate performs below level defined as A1.