Received: 28 October 2020 | Accepted: 23 October 2021

DOI: 10.1111/flan.12600

ORIGINAL ARTICLE

ACTFL Language Connects WILEY

Evaluating the reading and listening outcomes of beginning-level Duolingo courses

Xiangying Jiang[1] | **Joseph Rollinson**[1] | **Luke Plonsky**[2] | **Erin Gustafson**[1] | **Bozena Pajak**[1]

The Challenge

As more and more learners use digital apps to learn languages, it is important for the field of language learning to understand the effectiveness of these apps. This article presents the listening and reading proficiency of Duolingo learners when they reach the end of its beginning-level Spanish and French courses.

[1]Duolingo, Pittsburgh, Pennsylvania, USA

[2]Department of English, Northern Arizona University, Flagstaff, Arizona, USA

Correspondence
Xiangying Jiang, Duolingo, 5900 Penn Ave., Pittsburgh, PA 15206, USA.
Email: xiangying@duolingo.com

Abstract

Duolingo is a commercial language-teaching platform that offers free courses on the web and on mobile apps. This study reports the ACTFL listening and reading proficiency levels of adult Duolingo learners who had completed beginning-level courses in Spanish or French. The participants ($n = 225$) were learners residing in the United States, had little to no prior proficiency in the target language, and used Duolingo as their only learning tool. The Duolingo learners reached Intermediate Low in reading and Novice High in listening. No other skills were assessed. Their reading and listening scores were comparable with those of university students at the end of the fourth semester of study. The findings of the study suggest that Duolingo can be an effective tool for foreign language learning.

KEYWORDS

Duolingo, efficacy, foreign language, listening proficiency, reading proficiency

1 | INTRODUCTION

Online language courses, offered both by educational institutions and by commercial organizations, have seen accelerated growth in recent years. Some claim that online courses provide a valid alternative to face-to-face language classrooms, while many in the language education community remain skeptical (Lin & Warschauer, 2015) and demand "solid research studies in refereed sources comparing the language proficiency outcomes of online and face-to-face programs" (Tarone, 2015, p. 392).

This paper aims to address the existing gap in the literature by investigating the proficiency outcomes of the second language (L2) learners using Duolingo. Duolingo is a commercial language-teaching platform that offers free[1] online courses available on the web and on mobile apps. The specific goals of the present study were: (1) to evaluate—using standardized tests—the listening and reading proficiency levels of Duolingo users who learned with Duolingo only and completed the beginning-level Spanish or French courses, (2) to determine the amount of time they took to reach the end of the beginning-level course content, and (3) to compare their proficiency scores with those of university students in US-based language programs. Our findings contribute to an understanding of the kinds of target language gains that can be expected using one particular app-based program of study, Duolingo, both generally and as compared to the more familiar context of university-based language programs.

2 | LITERATURE REVIEW

2.1 | Online language learning and classroom language instruction

Most of the research on online language learning compares it with learning in a face-to-face classroom environment. Under the umbrella term of online language learning, there are a variety of instructional models and learning environments, which include web-facilitated classes, blended or hybrid courses, or fully online courses (R. Blake, 2011). Of course, there are also learners not affiliated with any formal courses. The overall findings of research on this topic demonstrate, among other benefits such as enhanced autonomy among learners and great adaptability to learners' needs, online language courses have comparable effects to face-to-face instruction, with no evidence of learners being disadvantaged even when compared on oral production assessments (R. J. Blake et al., 2008; Chenoweth & Murday, 2003; Chenoweth et al., 2006; Hampel & Hauck, 2004; Hampel, 2003; Isenberg, 2010; Lys, 2013; Sun, 2012; Ushida, 2005; Volle, 2005). In fact, the results reported in several meta-analyses show an advantage for technology-supported pedagogy. For example, Grgurović et al. (2013) showed a relatively small but consistent benefit of computer-assisted language learning or CALL-based instruction over face-to-face instruction for L2 development ($d = 0.26$ for studies demonstrating group equivalence at pretest). Notably, this difference held up across proficiency levels, educational levels, and target languages (for additional meta-analytic evidence of the advantages of technology in language instruction, see Plonsky & Ziegler, 2016).

This body of research, however, is limited in several respects. First, studies in this domain, like much of L2 research, have mainly focused on university contexts, which may not generalize to the broader population of adult language learners (Andringa & Godfroid, 2020; Plonsky, 2017). In addition, most studies have used researcher-made assessments and/or relied on more subjective outcome measures (e.g., perceptions and attitudes) instead of standardized

proficiency tests. Further limiting the generalizability of findings in this area, many studies have focused on specific technological tools such as an audio-graphic conferencing system (Hampel & Hauck, 2004; Hampel, 2003), Wimba Voice (Gleason & Suvorov, 2011), and voice blogs (Sun, 2012).

Among the various models of online language learning, this study focused on online courses offered by a commercial provider via language learning apps. As background to the present study, we review a number of other efficacy studies based on Duolingo and other commercially available apps. We also review studies on the proficiency outcomes of university language programs which serve as a source of comparison for language gains made by the participants in the present study.

2.2 | Effectiveness of commercial online language learning products

Due to the commercial nature of their products, companies sometimes hire researchers to carry out commissioned research. There is a noteworthy set of commissioned studies by Vesselinov and Grego across five online language learning products: Rosetta Stone (Vesselinov, 2009), Duolingo (Vesselinov & Grego, 2012), Babbel (Vesselinov & Grego, 2016a), Busuu (Vesselinov & Grego, 2016b), and Italki (Vesselinov & Grego, 2018). These research reports were published on company websites as white papers, not peer-reviewed journal articles. In these studies, the researchers followed a pretest–posttest design and investigated the effectiveness of the Spanish learning products of each company's product. The participants were non-Hispanic learners between the ages of 19–69, with a below-advanced Spanish proficiency. All five studies used the Web-based Computer Adaptive Placement Exam (WebCAPE, an adaptive exam that assesses vocabulary, reading, and grammar) as the primary data collection instrument and reported teaching effectiveness based on the points gained from the pretest to posttest and points gained per hour of study. Three studies (Vesselinov & Grego, 2016b, 2018; Vesselinov, 2009) also used the ACTFL Oral Proficiency Interview-computer version (OPIc) to assess participants' development in speaking ability. Overall, learners showed gains in WebCAPE points and some percentage of learners leveled up in ACTFL OPIc ratings. Due to the differences of pretest WebCAPE scores and OPIc levels, it is hard to compare the effectiveness across these products based only on gained points, gained points per hour of study, or percentage of learners leveling up. For the findings to be meaningful, this set of studies would have benefited from the more rigorous research designs, for example, control of prior proficiency, control of time on task, use of comparison groups, and use of more interpretable proficiency tests.

In two recent studies, Loewen and colleagues have also investigated the efficacy of online language learning products (Loewen et al., 2019, 2020). In a collaboration between two academics and a Babbel internal researcher, Loewen et al. (2020) examined the effectiveness of Babbel for learning Spanish. The study involved 54 participants who used Babbel to study Spanish for a minimum of 15 min/day during a period of approximately three months. The participants were college graduate and undergraduate students with an average age of 24 years and had an average of two classroom-based Spanish courses before the study. The study followed a pretest and posttest design based on measures of ACTFL OPIc, grammar, and vocabulary. The researchers found that after an average of approximately 12 h of learning on Babbel within 12 weeks, learners increased their oral proficiency by 0.7 ACTFL sublevels and made significant gains on grammar and vocabulary. The learning gains were associated with

the duration of time participants spent on Babbel and their overall level of interest in learning Spanish.

Loewen et al. (2019) is a case study on learning beginner-level Turkish with Duolingo. Unlike Loewen et al. (2020), the researchers of this study served as participants themselves. The researcher–participants were a professor and eight graduate students who were experienced language learners as well as researchers in language learning. They carried out the project to fulfill an obligatory class requirement. These true beginners of Turkish used Duolingo at least 1 h/week for 12 weeks. They were assessed with a summative achievement test which was used for a first-semester university-level Turkish class (Turkish 151 at the institution where the research was conducted). After an average of 29 h of learning Turkish on Duolingo, only one participant reached 70% of mastery on the Turkish 151 test. However, it is unclear whether the Turkish 151 test, designed for a particular university class, was appropriate as an outcome measure in the study. As an achievement test, the test might have strong content validity for the Turkish 151 class because it tested what had been taught in that class, but would not necessarily be appropriate to assess learning on Duolingo or any other program of instruction.

In contrast with the single-sample studies described thus far, some online language learning products have been compared with traditional classroom instruction and no evidence of disadvantage has been identified. Lord (2015, 2016) investigated the effectiveness of Rosetta Stone with data from 12 true beginners during a 16-week academic semester. The participants of the study were enrolled in a university beginning-level Spanish course. They were divided into three groups: a control group, a Rosetta Stone group, and a group that used Rosetta Stone materials as a course text in class, with four learners in each group. Two assessments were used at the end of the semester: the vocabulary and grammar portion of the Spanish College Level Examination Program (CLEP) test and the Versant Automated Oral Proficiency Test in Spanish. No significant differences were observed between the three groups on either measure, even though qualitative differences were noticed in the interview scripts favoring the control group. In addition to concerns related to the study's small sample, a substantial difference between groups was observed for time-on-task, with the control group averaging 109 h of learning and the Rosetta Stone group averaging only 48 h of learning.

In another recent study, Rachels and Rockinson-Szapkiw (2018) compared online language learning products with traditional classroom instruction. The authors employed a pretest-posttest design to compare face-to-face Spanish classroom instruction with Duolingo's Spanish course for English speakers in an elementary school. The participants of the study were 164 students from 11 third- and fourth-grade classes. Students from six classes used Duolingo to learn Spanish while the other five classes attended regular face-to-face Spanish classes. Both groups learned Spanish for 40 min/week for 12 weeks. Students were assessed on Spanish vocabulary and grammar with multiple-choice items. The same test was used in pretest and posttest. The researchers found no significant difference between the two groups and concluded that Duolingo was a useful tool for teaching Spanish to elementary students.

Several of the studies in this small set of studies provide some evidence of the effectiveness of online language learning products, indicating improvements in linguistic knowledge and no disadvantage compared to face-to-face learning. However, a few issues are noteworthy. First, there is a lack of involvement of independent researchers (see Lord, 2015, 2016, for a notable exception). The studies across different products were limited in the variety of authorship. For example, Vesselinov (and Grego) carried out commissioned studies on Rosetta Stone, Duolingo, Babbel, Busuu, and Italki. Loewen and colleagues have conducted studies on Babbel and Duolingo. The lack of research by academic scholars on commercial language learning

products has been observed by several researchers (e.g., Heift & Chapelle, 2012; Plonsky & Ziegler, 2016; Smith, 2017), who called for more participation of language learning researchers and educators in exploring the effectiveness of commercial products. Loewen et al. (2020) attributed the lack of scholarly interest to a number of reasons, including researchers' limited control when utilizing apps and their deterrence by the commercial nature of the apps. These reasons seem highly relevant and worthy of concern for the potential threat they present to the internal validity of this line of research. As the language learning field calls for rigorous research into the efficacy of commercial products, one way to address these concerns is to allow collaboration between external scholars and internal researchers, as in the study by Loewen et al. (2020), where university researchers and an internal Babbel researcher collaborated and co-authored the paper. The team involved in the present study, likewise, involves both industry- and university-based researchers. Even more trustworthy evidence might come from researchers who are completely independent of commercial entities.

Second, the outcome measures used in the studies were, in many cases, less than ideal. For example, as described above, Loewen et al. (2019) used a summative achievement test for a university class to assess learning on Duolingo; Vesselinov (and Grego) used a placement exam (WebCAPE) in all five studies they conducted. In the case of Vesselinov and Grego (2012, 2016a, 2016b, 2018), the researchers defined product efficacy as a gain of WebCAPE points per hour of study and provided estimates on the number of hours of study needed to be placed out of the first-semester university language course. Such findings were not only hard to interpret out of the immediate context, but can also be seen as making unwarranted claims. As a result, some scholars have expressed skepticism about some of the claims commercial language learning products have made about learner success, calling for more rigorous, research-based proficiency assessments (Tarone, 2015; van Deusen-Scholl, 2015).

2.3 | Proficiency outcomes of university language programs

As hubs of language learning, university-based language courses provide one possible source of comparison of the effectiveness of commercial online language learning products. Both settings have sought in recent years to move toward proficiency-based instruction and outcomes (e.g., Cox et al., 2018). In line with this movement, the Language Flagship Proficiency Initiative, supported by a grant from the National Security Education Program (Winke et al., 2014–2017), has funded the administration of proficiency assessments for language learners at the University of Utah, the University of Minnesota (Twin Cities), and Michigan State University. Students at varying semesters of undergraduate study (second to eighth semester) were assessed with the ACTFL Listening Proficiency Test (LPT), Reading Proficiency Test (RPT), and OPIc in 10 different languages with over 20,000 scores. Several publications have been available to professionals in language education related to the foreign language proficiency test data (Winke et al., 2014–2017) provided by the Flagship Proficiency Initiative (e.g., Rubio & Hacking, 2019; Tschirner, 2016). Considering the scope of the current study, the following review focuses on studies that reported the listening and reading proficiency levels of university students in Spanish and French.

Tschirner (2016) reported listening and reading proficiency levels at different milestones of undergraduate study based on data from more than 3000 participants learning seven languages at 21 institutions across the United States, although the majority of the test scores were from the foreign language proficiency test data (Winke et al., 2014–2017). More concretely, ACTFL

LPTs and RPTs were administered to first-, second-, third-, and fourth-year students from 2014 to 2015. Data were collected from learners of French, German, Japanese, Italian, Portuguese, Russian, and Spanish. The main findings were reported based on listening and reading proficiency levels in Spanish and French, which made up 82% of all tests completed. In both languages, there was a steady increase in proficiency levels over the semesters in both listening and reading, but listening proficiency levels were substantially lower than reading. By the end of the fourth semester, on average, students reached Intermediate Low (IL) in reading proficiency, but their listening proficiency was Novice Mid (NM), approaching Novice High (NH). Notably, the findings from Rubio and Hacking (2019), which reported findings from all three institutions of the Flagship Proficiency Initiative, were very consistent with those of Tschirner (2016): Among other results, after four semesters of instruction, reading reached IL in Spanish and French, but listening remained at NH.

Soneson and Tarone (2019) reported data from the Proficiency Assessment for Curricular Enhancement (PACE) project on ACTFL assessments of speaking, listening, and reading of seven languages at the University of Minnesota. Their findings reveal somewhat more rapid gains compared to Tschirner (2016) and Rubio and Hacking (2019). After two semesters of instruction, students in Spanish and French reached IL in reading, NH in listening, and IL in speaking. After four semesters of instruction, students reached Intermediate Mid (IM) in reading, IL in listening, and IM in speaking. The discrepancy between Soneson and Tarone (2019), on one hand, and the findings of other studies in the Flagship Proficiency Initiative, on the other, might be due to differences in instruction/exposure. According to Strawbridge et al. (2019), the PACE project was based on an enhanced curriculum that required five credit hours per semester, which was very likely more than language programs in other institutions that offer three or four credit hours per semester.

Similar to the studies described in this section thus far, Strawbridge et al. (2019) sought to track learners' speaking, listening, and reading proficiency ratings in French and Spanish in postsecondary programs. The researchers found that second- and fourth-semester students of both languages scored significantly lower in listening than in reading and speaking. At the end of the fourth semester, students in both languages reached IM in reading, IL in listening, and IM in speaking. However, as mentioned in relation to Soneson and Tarone (2019), the language programs under investigation offered five credit hours per semester for the first four semesters of language study.

Finally, Winke et al. (2020) provided a proficiency profile of the university undergraduate students in six languages (Arabic, Chinese, French, Portuguese, Russian, and Spanish) based on ACTFL speaking, listening, and reading proficiency data collected in Spring, 2017 of the Language Flagship Proficiency Initiative. Their findings largely resemble those of the other studies reviewed thus far: among students who are non-heritage speakers, fourth-semester French students achieved IL in both reading and listening; fourth-semester Spanish students reached IL in reading and NH in listening.

In sum, among the five studies reviewed in this section, Tschirner (2016), Rubio and Hacking (2019), and Winke et al. (2020) demonstrated that students at the end of the fourth semester of Spanish and French courses reached IL in reading proficiency and between NM and IL in listening proficiency, while Soneson and Tarone (2019) and Strawbridge et al. (2019) reported results of one ACTFL sublevel higher. The higher proficiency reported in these two studies seemed to be based on one language program which offers an enhanced curriculum. Across studies, four semesters of the study consisted of a range of two to five credit hours per semester for a total of eight to 20 credit hours total across semesters. These findings are summarized in Table 1.

Overall, the body of literature reviewed here can be summarized as follows. First, there is fairly strong evidence that technology-based instruction can be effective in fostering second language development. Such gains are especially robust when training occurs in the context of larger educational programs such as those offered by tertiary institutions. Evidence of the effectiveness of web-based language-learning apps also appears to be accumulating. However, as noted above, this line of investigation is somewhat limited not only by the number of investigations available to date but by certain study design features such as choice of outcome measures, sample sizes, and the lack of collaboration between independent (i.e., university-affiliated) and industry-based researchers. Finally, large-scale studies of university-based language learning provide a fairly clear picture of the range of proficiency-based outcomes at different levels of instruction. Considering the overlapping interests across these domains, of primary interest to the present study is to compare such outcomes with those of the users of commercially available apps to assess—using a standardized assessment—the effectiveness and efficiency of the latter.

2.4 | The current study

The current study aimed to shed light on the question of what proficiency outcomes Duolingo learners can expect to achieve. We did so by measuring the listening and reading proficiency levels of Duolingo learners who had completed the beginning-level material in the Spanish and French courses. (Follow-up studies will investigate the effectiveness of Duolingo for other skills such as speaking and writing). To better understand the proficiency levels learners have reached and the means to get there, user activity data such as time spent on learning were also analyzed. Finally, learners' proficiency levels as measured by standardized test scores were compared with the proficiency outcomes of students enrolled in US-based university language courses.

2.4.1 | Duolingo course structure

The beginning-level content of a Duolingo course includes five sections, each of which concludes with a "checkpoint" (see Figure 1, left). Sections consist of "skills," which are sets of lessons on a functionally coherent topic (e.g., Travel or School). There are a total of 114 skills in the beginning level of the Spanish course and 99 skills in the beginning level of the French course, as shown in Table 2. Each skill includes five difficulty levels with four to five lessons at

TABLE 1 Spanish and French reading and listening proficiency of fourth-semester university students

	Spanish		French	
Studies	**Reading**	**Listening**	**Reading**	**Listening**
Tschirner (2016)	IL	NM	IL	NM
Rubio and Hacking (2019)	IL	NH	IL	NH
Soneson and Tarone (2019)	IM	IL	IM	IL
Strawbridge et al. (2019)	IM	IL	IM	IL
Winke et al. (2020)	IL	NH	IL	IL

Abbreviations: IL, Intermediate Low; IM, Intermediate Mid; NH, Novice High; NM, Novice Mid.

each level, where the higher difficulty is achieved through exercises requiring progressively more recall and production. For example, the sentence-building exercise in Figure 1 (middle) is relatively easy compared to a similar exercise without a word bank. Learners are required to complete at least one difficulty level in each row to move on to the next row.

Duolingo uses a comprehension-based approach to foster long-term retention and to promote communication in the new language (for a review of evidence on the effects of comprehension-based instruction, see Shintani et al., 2013). The courses expose learners to vocabulary and grammar in sentences in the target language such that learners will gradually infer linguistic regularities from repeated exposure to and engagement with meaningful input. Furthermore, Duolingo lessons complement more implicit, comprehension-based learning with explicit feedback and explanations. For some structures, explicit explanation can offer a shortcut to more efficient learning. This is especially the case for features of the target language that may be difficult to notice from input alone (DeKeyser, 2003; Ellis, 2015). Duolingo courses also include longer-form, discourse-level content in the form of interactive story exercises (see Figure 1, right), which provide learners with opportunities to practice listening and reading skills. These exercises provide a real-world context for language use, demonstrate how language is organized beyond the sentence level, and feature more interactive and social aspects of the target language. Lessons of all types involve many opportunities for practice and repeated exposure to target language structures.

Duolingo courses are aligned to the Common European Framework of Reference (CEFR), an international standard for language proficiency (Council of Europe, 2001). The CEFR guides curricular development by focusing on communicative functions, that is, what learners actually are able to do with a language, such as asking for directions or ordering a cup of coffee.

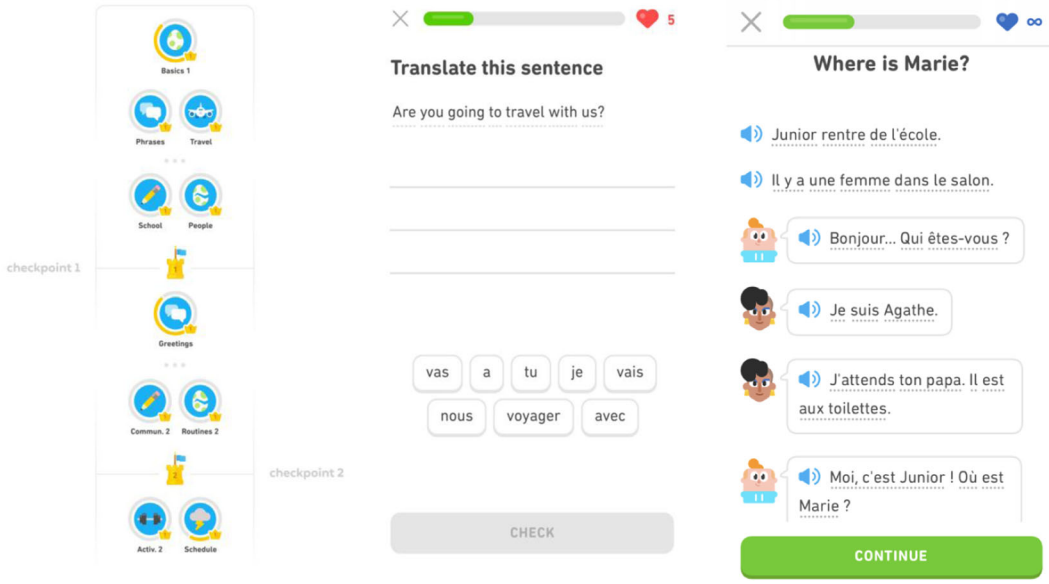

FIGURE 1 Example Duolingo course structure (left), example sentence-building exercise type (middle), and example story (right) [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Number of skills in each section of the Duolingo Spanish and French courses

Course section	Spanish: Number of skills	French: Number of skills
1	8	10
2	26	22
3	28	21
4	25	24
5	27	22
Total	114	99

2.4.2 | Research questions

As reviewed above, separate sets of studies have investigated the effectiveness of commercial online language learning products and university language programs. However, no direct comparisons have been made between proficiency outcomes of these two educational environments as the current study aims to do. The study also aimed to address some of the issues identified in the review of previous research on online language learning, such as collaboration between external academic scholars and internal researchers and the use of established proficiency measures. In particular, the current study investigated the following research questions:

1. What levels of reading and listening proficiency did Duolingo learners achieve upon reaching the end of the beginning-level Spanish and French courses? (RQ1)
2. What were the properties of learners' in-app activity—in terms of time spent studying, leveling up, and specific Duolingo features used—before reaching the end of the beginning-level course? (RQ2)
3. How did Duolingo learners' reading and listening proficiency scores compare with proficiency outcomes of US-based university students in Spanish and French courses based on ACTFL reading and listening proficiency tests? (RQ3)

3 | METHODS

3.1 | Participants

The participants of the current study were 135 Spanish learners and 90 French learners using the Duolingo product. They were learners who: (1) were at least 18 years old; (2) had an IP Address in the United States; (3) had self-reported no or little prior proficiency in the target language; (4) reached the end of Section 5 of the course; (5) reported using Duolingo as the only tool to learn the target language;[2] and 6) had proper computer equipment for online testing (see further information on the recruitment procedures below).

A combination of program-recorded data and response to background survey questions was used to select participants who met all these criteria and who voluntarily and independently chose to use Duolingo to learn French or Spanish. With regard to self-reported no or little prior proficiency in the target language, only learners who reported prior proficiency of 0–2 on a 0–10 scale were included,

with 0 meaning "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 when they reach the first checkpoint for the purposes of learner analytics and not for course placement.

Demographic and other background information were also collected through the survey. Some general characteristics of the participants include the following: Among 210 participants who reported age, it ranged from 18 to 83 with a mean of 43.99 (SD = 15.54). In terms of gender, 49% of the participants identified themselves as male and 48% as female. Seventy-eight percent of the participants listed their ethnicity as Caucasian, 13% as Asian, and 3% as African American. Thirty-nine percent of the participants reported having a bachelor's degree as their highest level of education, 37% having a master's degree, and 14% having a doctoral degree. Finally, 74% of the participants reported speaking only English before age 6; 8% were early bilingual speakers of English and another language; and 18% of the participants did not speak English before age 6 (their first languages varied widely and none of them were heritage speakers of the target language). For a more detailed by-course description of participant background information, see Appendix A.

3.2 | Instruments

3.2.1 | ACTFL LPT and RPT

The ACTFL LPT and RPT were used as the main data collection instruments. The ACTFL LPT and RPT are standardized tests for the global assessment of reading and listening ability (ACTFL, 2013, 2014). They measure how well test-takers spontaneously comprehend the texts and discourse they read or listen to as described in the ACTFL 2012 Proficiency Guidelines. ACTFL has 10 levels in its proficiency rating scale, from low to high in the order of Novice (low, mid, high), Intermediate (low, mid, high), Advanced (low, mid, high), and Superior. For the purpose of this project, Form E of the tests was used, which targets proficiency levels between Novice Low and Advanced Low. The tests, paid for by Duolingo, were administered to each participant online by a remote human proctor employed by ACTFL/Language Testing International. The participant was asked to read or listen to 15 passages and answer three multiple-choice questions after each passage. Each test was given an ACTFL rating immediately after the test was submitted. ACTFL ratings were coded numerically by following the 1–10 point scale as in previous studies (e.g., Rubio & Hacking, 2019; Tschirner, 2016; Winke et al., 2020). See Table 3 for the mapping between the point scale and each proficiency sublevel.

3.2.2 | Background survey

The questionnaire included sets of questions related to language background, demographic information, self-assessment of proficiency development, feedback about the Duolingo product, and a set of questions for participant selection mentioned earlier. The questionnaire can be found in Appendix B.

3.3 | Data collection procedures

Data collection took place during May–July 2020. Learners with an IP Address in the United States and a prior proficiency of 0–2 were contacted with an e-mail when they reached the end of Section 5

TABLE 3 ACTFL ratings and numerical coding

ACTFL level	ACTFL rating	Abbreviation	Numerical coding
Novice	Novice Low	NL	1
	Novice Mid	NM	2
	Novice High	NH	3
Intermediate	Intermediate Low	IL	4
	Intermediate Mid	IM	5
	Intermediate High	IH	6
Advanced	Advanced Low	AL	7
	Advanced Mid	AM	8
	Advanced High	AH	9
Superior	Superior	S	10

TABLE 4 Numbers in the data collection funnel

	Number of e-mails sent	Number of surveys submitted	Number of qualified learners	Number of participants
Spanish	2175	296	171	135
French	1038	193	116	90

of the Spanish or French Duolingo course. In the e-mail, they were invited to participate in a research study and were encouraged to submit the background survey. They were selected to participate in the study if their responses indicated that (1) they did not take classes or use other programs/apps to learn the target language during the period of learning on Duolingo and (2) they had access to proper equipment for taking the test.

Participants completed one ACTFL proficiency test at a time, with the order of tests (reading and listening) randomized across participants. Each time a test was ordered, the participant received an e-mail from Language Testing International (LTI) with their test ID and instructions about how to schedule a time for the test. After they finished the first test, the second test was ordered for them and they were again contacted by LTI to take the second test. They went through the same process to schedule and take the test. Each participant received $100 from Duolingo after completing both tests. Table 4 shows the funnel for data collection.

A few participants did not take both tests. Among a total of 135 Spanish-learner participants, 132 reading and 131 listening scores were collected. Among a total of 90 French-learner participants, 88 reading, and 89 listening scores were collected.

3.4 | Analyses

Descriptive statistics were calculated to answer the first and second research questions on the proficiency outcomes of Duolingo learners and their in-app activity until reaching the end of the beginning-level content. For the third research question on the comparison of proficiency outcomes

between university students and Duolingo learners, *t* tests were carried out for each language skill with the R statistical package (R Core Team, 2020).

4 | RESULTS

4.1 | Proficiency outcomes of Duolingo Learners

The reading and listening proficiency ratings of Duolingo learners who participated in the current study are presented in Figure 2. The ratings in Spanish reading, French reading, and French listening were normally distributed; however, the ratings in Spanish listening were positively skewed. Two-thirds of the Spanish listening proficiency ratings were at the Novice level.

On the basis of the numerical coding of the proficiency ratings on a 1–10 point scale presented in Table 3 above, Table 5 presents the summary data with mean scores and standard deviations. Overall, Spanish and French reading scores were between IL (4) and IM (5), while listening scores were at least one level below reading scores. Spanish listening was approaching NH and French listening was at NH.

4.2 | In-app activity of Duolingo learners

The reading and listening proficiency scores demonstrated the extent of target language development that occurred in the beginning-level Duolingo Spanish and French courses; however, another aspect of efficacy is how efficient the learning process is. To understand the degree of efficiency of the Duolingo Spanish and French courses, the amount of time Duolingo learners took to reach the end of the beginning-level course content was calculated. The total number of hours that the study participants spent in all Duolingo sessions in the given course were computed and summarized in Figure 3. This calculation is documented in Appendix C. The mean number of hours that learners across the two courses spent studying on Duolingo was 141 (median: 112). French learners spent on average about 20 h less than the Spanish learners to finish the beginning-level course, which is likely due to fewer course skills in French, as reported in Table 2 above. Learners also varied considerably in the number of days elapsed between their first lesson on Duolingo for the target language and participation in the current study; on average, 562 days passed for Spanish learners (median = 412 days, SD = 551 days) and 634 days passed for French learners (median = 359 days, SD = 707 days). The number of days in which the learners used the app during these periods, however, varies immensely across the sample.

As expected, a high degree of variation exists in the amount of time learners spent learning on Duolingo (Spanish: SD = 118, IQR = [44–213]; French: SD = 115, IQR = [39–192]). Due, at least in part, to this variation, very small and nonsignificant correlations (Spearman's ρ) between time spent using Duolingo and test scores for either Duolingo course were observed (see Figure 4). Variation in time spent learning on Duolingo is expected due to low minimum requirements to progress through sections of the course. While each course skill has five difficulty levels, learners were required to complete only one of those levels to move on to the next row. Some learners reached the fifth difficulty level in all skills while others did the minimum to move along the course, thus leading to large between-participant differences in the number of hours spent learning on Duolingo. Furthermore, this time spent learning measure potentially spans many years of study; some participants may have completed fewer hours more recently, while others completed many hours spanning several

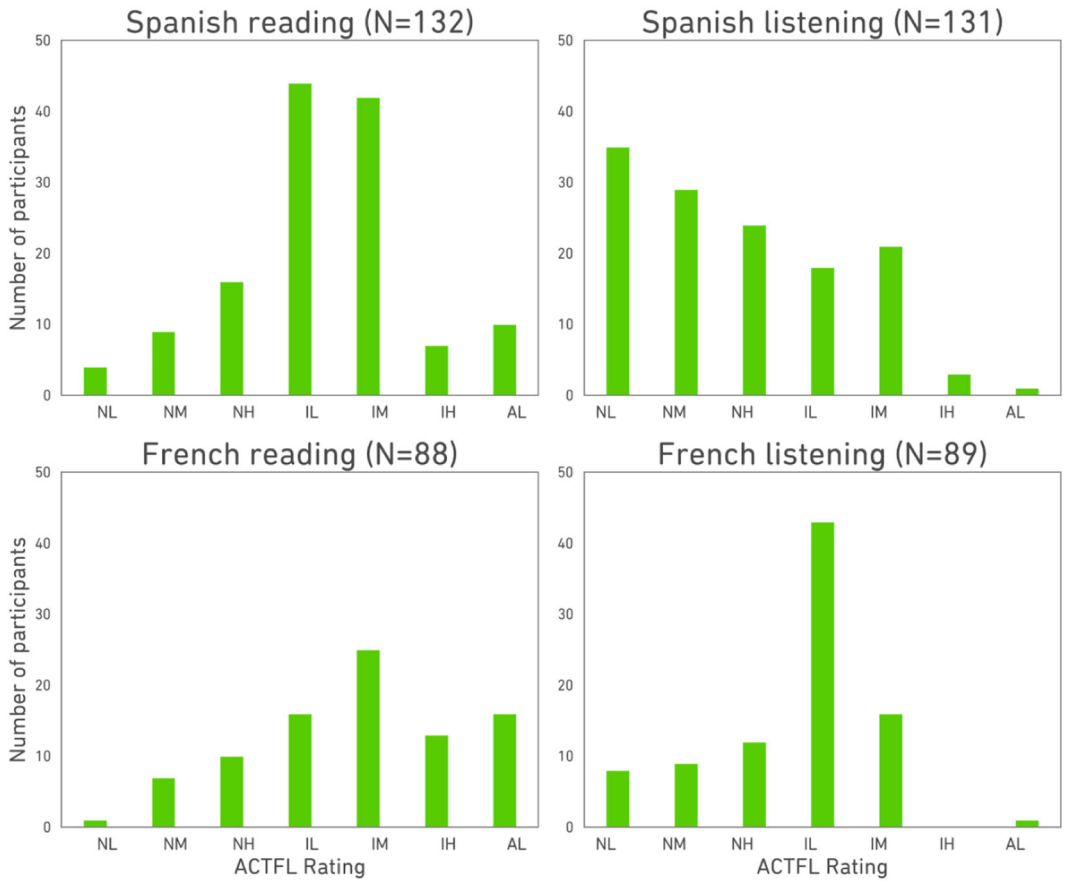

FIGURE 2 Distribution of ACTFL proficiency ratings of Duolingo learners. The *x*-axis shows ACTFL rating acronyms (see Table 3) [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 5 Spanish and French reading and listening scores of Duolingo learners

	N	Mean/median score (SD)	ACTFL rating
Spanish reading	132	4.30/4.0 (1.34)	Intermediate Low
Spanish listening	131	2.80/3.0 (1.54)	approaching Novice High
French reading	88	4.82/5.0 (1.55)	approaching Intermediate Mid
French listening	89	3.61/4.0 (1.22)	Novice High

years. These differences in participant behavior—and the resulting variation in the time spent learning measure—make it difficult to draw conclusions about the relationship between total time spent learning on Duolingo and learning outcomes measured by the ACTFL assessment. Future studies could address these issues and provide stronger signals about this relationship by using a pre- and posttest design with more control over the time spent learning over the course of the study.

On the days the learners chose to study (restricted to days between starting and completing Section 5, the final section before qualifying for study participation), they completed around eight lessons on average (Spanish: mean = 8.5, median = 6.9; French: mean = 7.9, median = 6.3).

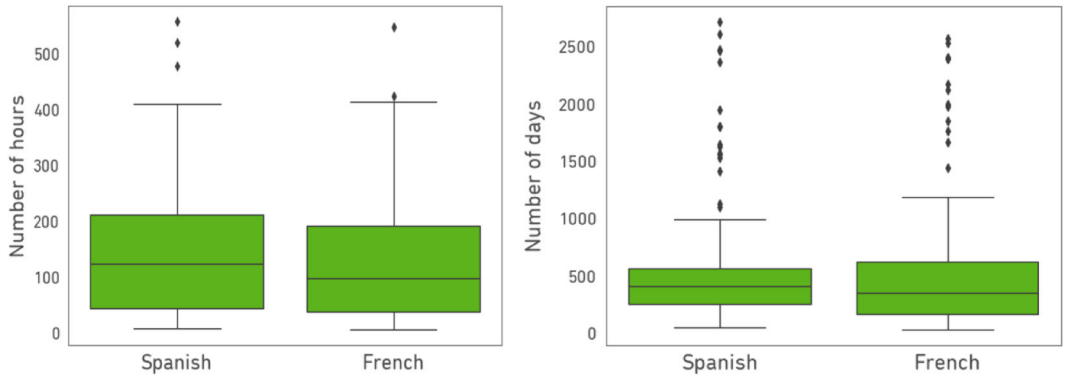

FIGURE 3 Distributions of hours spent to complete Duolingo Sections 1–5 (left) and number of days elapsed between the first lesson in target language and study recruitment (right) [Color figure can be viewed at wileyonlinelibrary.com]

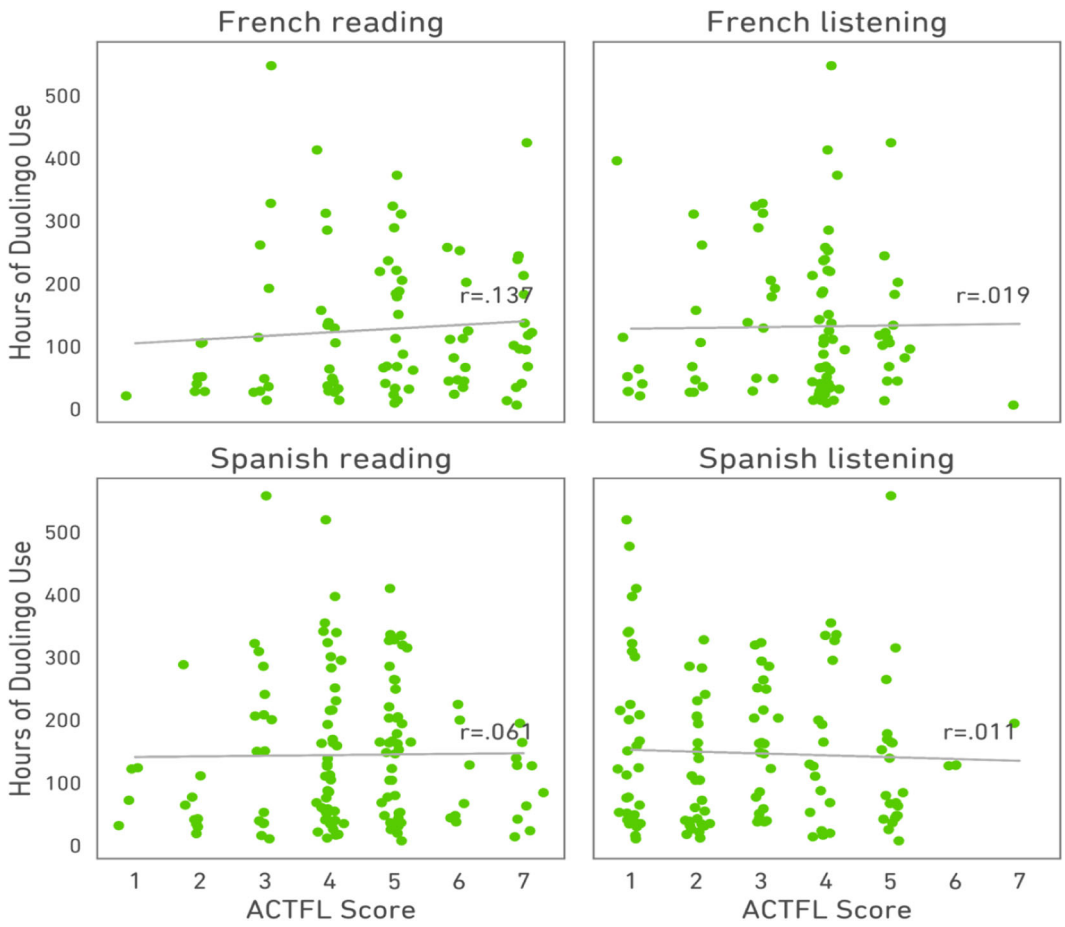

FIGURE 4 Correlation (Spearman's ρ) between hours of Duolingo use and ACTFL scores [Color figure can be viewed at wileyonlinelibrary.com]

However, as with overall time spent learning, considerable variation was observed here (Spanish: SD = 7.0; French: SD = 5.1; see Figure 5 for full distribution). Learners also varied in the number of days taken to complete Section 5 (Spanish: mean = 81.2, median = 64.5, SD = 68.9; French: mean = 90.7, median = 75.5, SD = 62.6; see Figure 5 for full distribution).

As noted in the Introduction, Duolingo courses are broken into "skills" that target certain vocabulary and/or grammatical concepts. Each skill includes five difficulty levels and learners are required to complete all skills in a given section at the lowest difficulty level (Level 0) before they can complete the Checkpoint for that section. Aside from this requirement for progressing to new sections, Duolingo learners are free to study however they want; some learners choose to focus on exploring new content (e.g., Level 0 lessons) while others study up on more familiar content (e.g., Level 1+ lessons). Due to this freedom, considerable variation was seen in the types of lessons that participants in this study completed (Figure 6).

The distributions in Figure 6 shows that for many participants, the majority of lessons completed are Level 0; for more than 20% of participants (Spanish: 22.9%; French: 21.3%), Level 0 comprised at least half of all lessons completed. Other participants spent more time "leveling up" by studying with more difficult exercises. Bimodal distributions for Levels 2–4 sessions were observed, which means some users rarely "leveled up" (i.e., those focusing mainly on Level 0) and others spent time completing higher-level lessons.

Outside of standard lessons, many participants also completed Stories, which provide learners with discourse-level listening and reading practice. On average, 8%–10% of lessons completed by the participants were Stories (Spanish: 8.5%; French: 9.9%). Participants spent relatively little time completing "practice" lessons; on average, fewer than 4% of lessons completed were either practice type (Spanish: 3.8%; French: 3.7%)

4.3 | Comparison with university courses

The third research question of this study was to compare the proficiency outcomes of Duolingo learners with the outcomes of US-based university students in language courses provided by the foreign language proficiency test data (Winke et al., 2014–2017). Although the learner

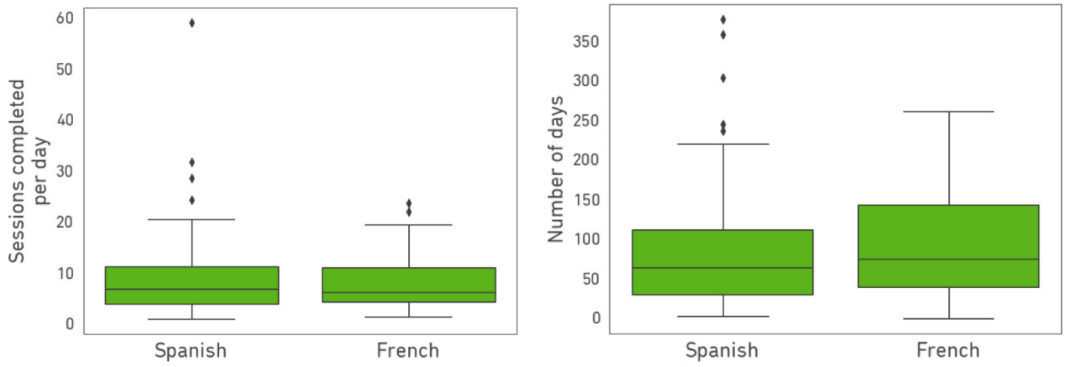

FIGURE 5 Distribution of Duolingo lessons completed per day (left) and number of days taken to complete Section 5 (final course section before ACTFL testing; right) [Color figure can be viewed at wileyonlinelibrary.com]

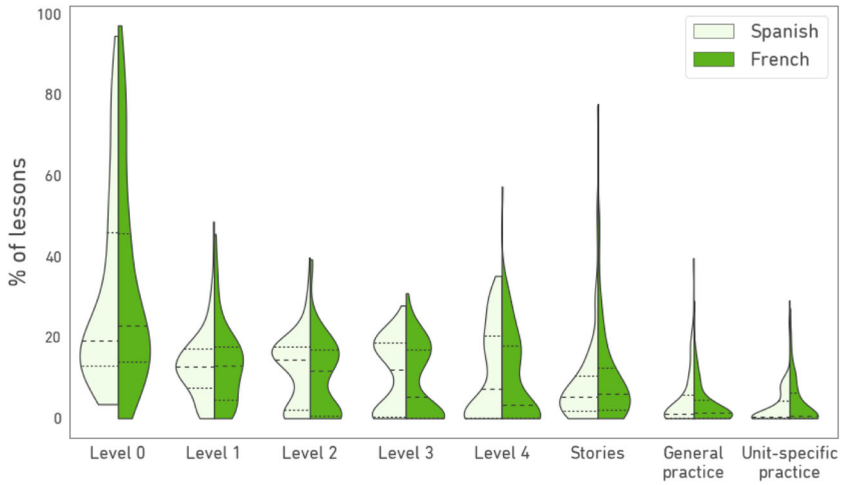

FIGURE 6 Distribution of Duolingo lesson types completed between July 2018 and study qualification, as proportion of all lessons completed by a given participant. Distribution for each lesson type for a given course is shown with a density plot [Color figure can be viewed at wileyonlinelibrary.com]

populations may be vastly different, with a much greater homogeneity in ages among the university-based learners, it is informative to establish correspondences between learner proficiency outcomes across distinct educational environments.

As mentioned earlier, the foreign language proficiency test data (Winke et al., 2014–2017) include assessments at various semesters of undergraduate study. The Spanish data include scores from second to eighth semester (except for the seventh semester) and the French data include scores from second to eighth semester (except for the fifth semester). On the basis of the performance of Duolingo learners reported above (see Table 5), a statistical comparison between Duolingo learners and fourth-semester university students was conducted. The fourth semester is the highest level in most university basic language programs before the traditional—if somewhat antiquated—"bridging" occurs into courses for language majors and minors (Graman, 1997) and is often used as the criterion for meeting degree- or university-based language requirements. The ACTFL ratings in the university data were coded numerically in the same way as the Duolingo data based on Table 3. Table 6 summarizes the descriptive statistics as well as the results of a series of t tests comparing the Duolingo and university learner performance.

To assess whether there were significant differences between Duolingo learners and university fourth-semester students, separate Welch two-sample t tests on each of the four sets of scores were carried out. No significant differences and small effect sizes (d; see Plonsky & Oswald, 2014) were found on Spanish listening ($t = -1.74$, $p > .05$, Cohen's $d = -0.24$), Spanish reading ($t = 0.35$, $p > .05$, Cohen's $d = 0.04$), and French listening ($t = 1.41$, $p > .05$, Cohen's $d = 0.21$), which suggests that Duolingo learners were not significantly different compared with university students at the end of their fourth semester. A significant and moderately sized difference for French reading was found ($t = 4.36$, $p < .05$, Cohen's $d = 0.72$), which showed that Duolingo learners performed significantly better than university students at the end of their fourth semester.

To show how Duolingo proficiency scores align with semester-based university data, second- to sixth-semester data from US university students were included in Figure 7. Please note that fifth-semester French data were not available in the university data set.

TABLE 6 Comparisons between Duolingo participants and fourth-semester US-based university students on Spanish and French reading and listening scores

Language skill	Groups	*N*	Mean (SD)
Spanish listening	Duolingo	131	2.80 (1.54)
	Universities	774	3.05 (1.49)
Spanish reading	Duolingo	132	4.30 (1.34)
	Universities	782	4.26 (1.59)
French listening	Duolingo	89	3.61 (1.22)
	Universities	422	3.39 (1.62)
French reading	Duolingo	88	4.82 (1.55)
	Universities	424	4.03 (1.55)

5 | DISCUSSION AND CONCLUSION

5.1 | Summary of findings

This study assessed the reading and listening proficiency of Duolingo learners who had completed the beginning-level material in the Spanish and French courses, analyzed their in-app activities, and compared their proficiency scores to those of fourth-semester university students on the same measures. The study aimed to answer three research questions. The first question asked about the levels of reading and listening proficiency that Duolingo learners achieved upon reaching the end of the beginning-level Spanish and French courses. Complementary to RQ1, our second research question was concerned with learners' in-app activity such as time spent studying, leveling up, and the specific Duolingo features they used en route to reaching the end of the beginning-level course. RQ3 inquired about how Duolingo learners' reading and listening proficiency scores compare with the proficiency outcomes of US-based university students in Spanish and French courses.

To answer the first research question, the results indicated that Duolingo learners who had completed the beginning-level material in Spanish reached IL in reading (according to ACTFL RPT) and approached NH in listening (according to ACTFL LPT), while learners studying French approached IM in reading and reached NH in listening. The current study was designed to address limitations in previous efficacy research for online language learning platforms, such as insufficient involvement from independent researchers and lack of rigor in the instruments used to measure proficiency. Due to study design and instrument differences for other research on the efficacy of online language learning platforms, comparison to these studies is difficult. However, the current results demonstrate the ability of online language learning platforms to teach to intermediate-level proficiency in reading and advanced novice-level in listening. Future studies will assess learner proficiency in other core competencies, such as speaking and writing, and investigate additional gains afforded by more advanced content.

Language learning apps are thought to be good for developing decontextualized linguistic knowledge and Duolingo is considered one example of such apps (Krashen, 2014). Although beginning-level Duolingo lessons focus on vocabulary and grammar at the sentence level, the findings demonstrated that learners were able to transfer discrete linguistic knowledge to

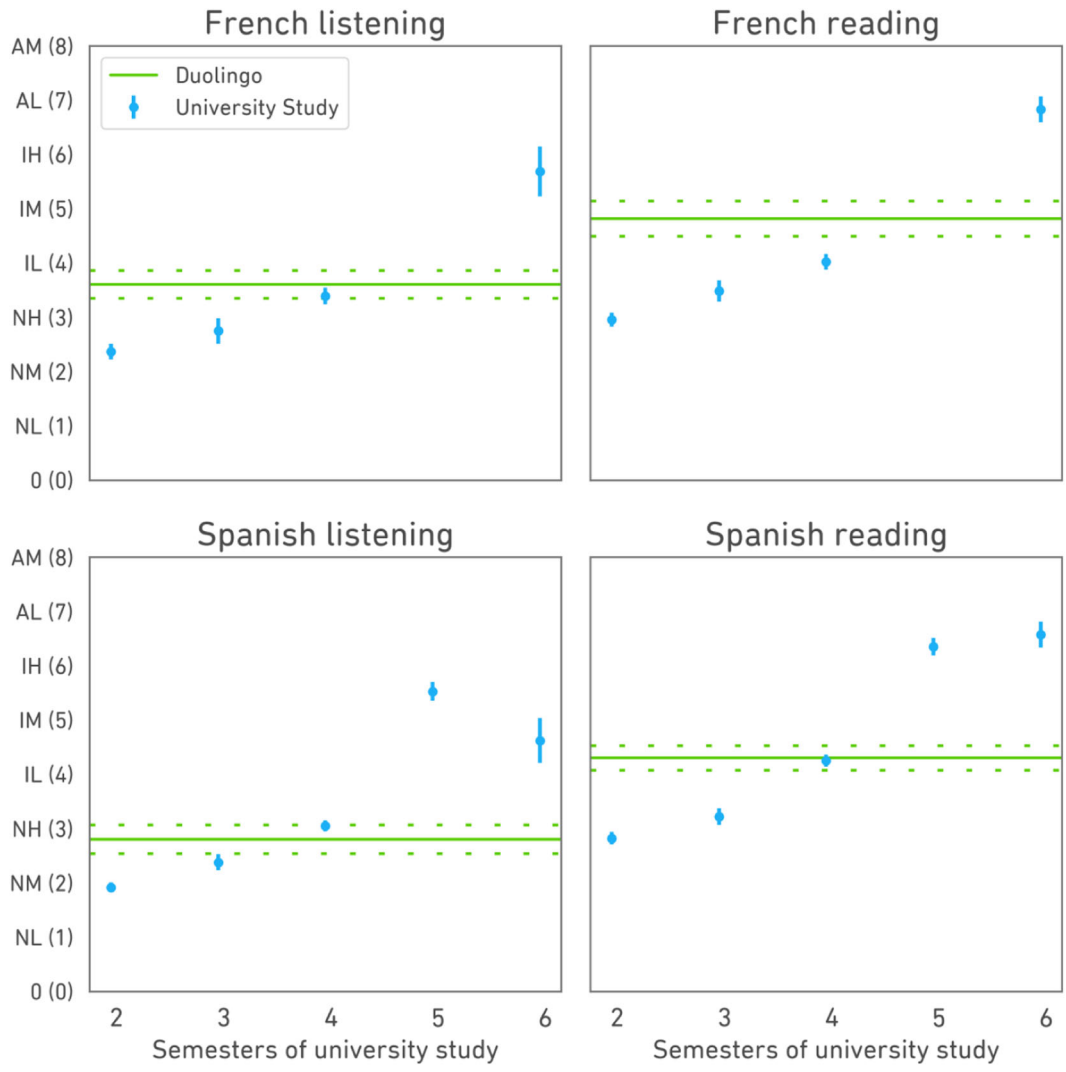

FIGURE 7 Comparison of mean ACTFL proficiency test scores for Duolingo and the university study, with 95% confidence intervals. See Table 3 for the proficiency ratings shown on the *y*-axis [Color figure can be viewed at wileyonlinelibrary.com]

integrative tasks such as reading and listening comprehension. This type of knowledge transfer and integration was also evidenced in Loewen et al. (2020), which shows that even with limited opportunities for oral production on Babbel, the explicit vocabulary and grammar knowledge that Babbel learners mastered led to encouraging gains in oral proficiency. The transfer of explicit linguistic knowledge is supported by Skill Acquisition Theory (e.g., DeKeyser, 2015), which states that practice and repetition can lead to proceduralization of explicit knowledge and hence improved language learning outcomes. With such findings in mind, Loewen et al. (2020, p. 19) proposed that the field of second-language acquisition should "recognize the pedagogical potential of widely used modern apps" and "abandon earlier characterizations of language learning apps as merely 'mechanical practice of selected and graded grammatical

phenomena... in the form of drills'" by citing Heift and Vyatkina (2017). The authors of this study concur with Loewen et al. on this proposal. However, of course, any claims of skills transfer among Duolingo learners would need to be tested empirically.

Indeed, the rate of development does not seem to be the same for all language skills and we would emphasize, again, that the present study only reports gains made in the two receptive skills of reading and listening. Although the participants' reading and listening scores were moderately correlated, the listening proficiency of Duolingo learners was significantly lower compared to reading proficiency, which replicated the findings of Tschirner (2016) and Rubio and Hacking (2019) for university students. Although both listening comprehension and reading comprehension are receptive skills, the comprehension processes have been found to be mostly modality-specific (Wolf et al. 2019). For learners at early stages of language learning, listening comprehension demands a higher level of attention, exerts a heavier load on working memory, and requires the ability for speedy decoding and processing of transient audio input (see, e.g., Bloomfield et al., 2010; Wallace, 2020). In contrast, learners' decoding process in reading is facilitated by the availability of visually presented text (Spoden et al., 2020; Vandergrift & Baker, 2015). As a result, listening comprehension is often more challenging than reading comprehension for second language learners. Some researchers also attributed students' lower listening proficiency to insufficient attention to auditory input and exercises in classroom instruction and called for more emphasis on listening development in instructional practices (Tschirner, 2016).

Analysis for RQ2 demonstrated that the median amount of time that the participants took to complete the beginning-level material was 112 h (99 h for French learners and 125 h for Spanish learners). On the days that the participants chose to study content in the beginning-level course section, they completed eight lessons on average, with the majority of the lessons at Level 0, which is the lowest and required level to progress through the course. Substantial variation in time spent studying may explain, at least in part, the lack of correlation between assessment outcomes and total time spent, a finding that contrasts with those of Loewen et al. (2020). The self-directed nature of the Duolingo learning platform contributes to this variation and complicates our interpretation of how the amount of learning effort contributes to assessment outcomes; for example, we observed bimodal "leveling up" behavior, where some learners choose to complete more difficult skill levels while others rarely do. It appears that the quality of the time spent in terms of activities, lessons, and attention given, may matter just as much as the quantity of time spent using the app. Future studies could address these issues and provide stronger signals about this relationship by using a pre- and posttest design, which allows for more control over the time spent learning over the course of the study. Other studies have had success with this design (e.g., Loewen et al., 2020).

In comparing listening and reading proficiency between Duolingo learners and university students in language classes, the results indicated that the proficiency scores of Duolingo learners aligned with those of fourth-semester university students. Specifically, when Duolingo Spanish and French learners reached Checkpoint 5 at the end of the beginning-level course content on Duolingo, their Spanish reading, Spanish listening, and French listening proficiencies were comparable to what university students accomplished in four semesters of classes, while their French reading proficiency was significantly higher than fourth-semester university students. Previous studies have also compared proficiency following the use of online language learning products to university classroom outcomes. Lord (2015, 2016) found similar levels of achievement for classroom learners compared to learners using only Rosetta Stone over the course of a semester. Similarly, Rachels and Rockinson-Szapkiw (2018) observed no significant differences between outcomes for third and fourth graders using Duolingo to learn Spanish and those who received classroom instruction. The findings of the current study—combined with those from previous research—provide evidence that online language

learning products can be effective methods for learning an additional language, at least in reading and listening.

5.2 | Limitations and directions for future research

The findings of the current study do not represent the overall effectiveness of Duolingo or university language courses, so they should not be overgeneralized. Participants of the study were only compared on reading and listening skills while teaching effectiveness can be reflected in other skills and abilities. In addition, there were a number of differences between the participants of the study and the university student sample. The university proficiency project tested full-time university students from a more homogeneous age range, while the participants of the current study were more varied demographically and included mostly post-university older adults. Similarly, the participants' motivations for language learning could also be more varied than university students, who included both those studying to meet a requirement and some who would later declare majors or minors in the language. These differences may put into question the comparability of the learners and the learning that took place in these two very different settings. The availability of the university proficiency data made this comparison possible; however, the comparison between Duolingo learners and university students should not be interpreted as competition between online language learning apps and university language programs. The aim in comparing learning outcomes from the two contexts is, rather, as a means to benchmark the progress made by Duolingo learners relative to a more familiar and traditional setting.

The current study tested learners when they reached Checkpoint 5 independently. For future research, treatment studies with a pre- and posttest design will allow more control of learning time and participant factors that were self-reported in the present study, including prior proficiency, exposure to the target language outside of Duolingo, and the exclusion of other learning tools. This study focused on listening and reading proficiency, which are both receptive skills. Learners were not assessed in speaking (as in Rubio & Hacking, 2019) or writing. In subsequent studies, Duolingo's effectiveness in developing learners' productive skills will be evaluated as well. Doing so will provide a better understanding of whether and to what extent Duolingo learners' success in receptive skills generalizes to other skills.

5.3 | Pedagogical implications

The study indicates that using Duolingo as a tool to develop reading and listening proficiency may be at least as effective as developing these proficiencies in a university classroom through traditional pedagogies. Although Duolingo courses mostly teach vocabulary and grammar at the sentence level (with some longer-form content available in the form of short stories and podcasts), the results of this study also suggest that the seemingly discrete vocabulary and grammar knowledge can be applied to integrative tasks such as listening and reading comprehension.

The findings of the study indicate that learners who use Duolingo as a tool for the self-directed study show substantial proficiency development. As we might expect, the usage data from the present study indicates a very slightly positive relationship between learners' total hours spent using the app and their reading and listening scores. In other words, more time on the app is associated with greater gains. However, Duolingo app usage data also points to vast variability in the time (hours) and intensity of learning that participants took to complete the

first five sections of their course. Consequently, it would be premature to make any suggestions regarding when and how the app might be used to maximize its efficiency. However, we plan to address this question in a future study.

In addition to self-directed learners, classroom teachers have used Duolingo to their advantage and benefited their students (Munday, 2016, 2017), suggesting that the app is also a useful tool to complement other types of language instruction. For instance, if vocabulary and grammar practice can be largely done by students as homework using apps such as Duolingo, more class time can be directed toward the teaching of culture and other communicative skills.

5.4 | Conclusion

This study assessed the reading and listening proficiency outcomes of Duolingo learners who had little to no prior knowledge of the target language and used Duolingo as the only learning tool. The findings demonstrated that learners who finished the beginning section of the Duolingo Spanish or French course reached IL in reading proficiency and NH in listening proficiency. These proficiency scores of Duolingo learners were comparable with the proficiency outcomes of students at the end of the fourth semester in university-based language programs (Rubio & Hacking, 2019; Tschirner, 2016). In conducting this study, we hope to have shed light on the potential effectiveness and comparability of Duolingo, as measured through standardized tests, to more traditional settings. Future studies will continue to build on our findings at other levels of study, in other linguistic domains, and in other target languages.

ENDNOTES

[1] Duolingo offers all learners free access to the entirety of its instructional materials. Learners can optionally purchase a subscription, Duolingo Plus, but the subscription does not give access to any additional educational content. Instead, Duolingo Plus offers an ad-free experience, the ability to download lessons for offline use, and other gamification features.

[2] As noted by an anonymous reviewer, although we screened participants for exposure to instructional materials other than Duolingo, we did not inquire about noninstructional (i.e., work or socially oriented) exposure to the target language. However, participants were only recruited if they had an IP address within the United States. Further, we would note that the same types of exposure may have also been present among the university-based learners who serve as a point of comparison for the present study.

REFERENCES

ACTFL. (2013). ACTFL reading proficiency test (RPT). Familiarization manual and ACTFL proficiency guidelines 2012—reading. Retrieved on July 2, 2020, from https://www.languagetesting.com/pub/media/wysiwyg/manuals/ACTFL_FamManual_Reading_2019.pdf

ACTFL. (2014). ACTFL listening proficiency test (LPT). Familiarization manual and ACTFL proficiency guidelines 2012—listening. Retrieved on July 2, 2020, from https://www.languagetesting.com/pub/media/wysiwyg/manuals/ACTFL_FamManual_Listening_2019.pdf

Andringa, S., & Godfroid, A. (2020). Sampling bias and the problem of generalizability in applied linguistics. *Annual Review of Applied Linguistics*, *40*, 134–142.

Blake, R. (2011). Current trends in online language learning. *Annual Review of Applied Linguistics*, *31*, 19–35.

Blake, R. J., Wilson, N. L., Cetto, M., & Pardo–Ballester, C. (2008). Measuring oral proficiency in distance, face-to-face, and blended classrooms. *Language Learning & Technology*, *12*, 114–127.

Bloomfield, A., Wayland, S. C., Rhoades, E., Blodgett, A., Linck, J., & Ross, S. (2010). *What makes listening difficult: Factors affecting second language listening comprehension* (Technical Report no.: TTO 81434 E. 3.1). Center for Advanced Study of Language (CASL), University of Maryland.

Carpenter, S. K. (2020). Distributed practice/spacing effect. In L.-F. Zhang (Ed.), *Oxford research encyclopedia of education*. Oxford University Press.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354–380.

Chenoweth, N. A., & Murday, K. (2003). Measuring student learning in an online French course. *CALICO Journal*, *20*, 284–314.

Chenoweth, N. A., Ushida, E., & Murday, K. (2006). Student learning in hybrid French and Spanish courses: An overview of language online. *CALICO Journal*, *24*, 115–146.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.

Cox, T. L., Malone, M. E., & Winke, P. M. (2018). Future directions in assessment: Influences of standards and implications for language learning. *Foreign Language Annals*, *51*, 104–115.

DeKeyser, R. (2003). Implicit and explicit learning. In C. Doughty, & M. Long (Eds.), *Handbook of second language acquisition* (pp. 313–348). Blackwell.

DeKeyser, R. M. (2015). Skill acquisition theory. In J. Williams, & B. VanPatten (Eds.), *Theories in second language acquisition: An introduction* (2nd ed., pp. 94–112). Routledge.

Dewaele, J.-M., Witney, J., Saito, K., & Dewaele, L. (2018). Foreign language enjoyment and anxiety: The effect of teacher and learner variables. *Language Teaching Research*, *22*, 676–697. https://doi.org/10.1177/1362168817692161

Ellis, N. C. (2015). Implicit AND explicit language learning: Their dynamic interface and complexity. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages*. Benjamins.

Gleason, J., & Suvorov, R. (2011). Learner perceptions of asynchronous oral computer-mediated communication tasks using Wimba Voice for developing their L2 oral proficiency. In S. Huffman, & V. Hegelheimer (Eds.), *The role of CALL in hybrid and online language courses*. Iowa State University.

Graman, T. L. (1997). The gap between lower- and upper-division Spanish courses: A barrier to coming up through the ranks. *Hispania*, *70*, 929–935.

Grgurović, M., Chapelle, C. A., & Shelley, M. C. (2013). A meta-analysis of effectiveness studies on computer technology-supported language learning. *ReCALL*, *25*, 165–198.

Hampel, R. (2003). Theoretical perspectives and new practices in audio–graphic conferencing for language learning. *ReCaLL*, *15*, 21–36.

Hampel, R., & Hauck, M. (2004). Towards an effective use of audio conferencing in distance language courses. *Language Learning & Technology*, *8*, 66–82.

Heift, T., & Chapelle, C. A. (2012). Language learning through technology. In A. Mackey, & S. Gass (Eds.), *The Routledge handbook of second language acquisition* (pp. 555–569). Routledge.

Heift, T., & Vyatkina, N. (2017). Technologies for teaching and learning L2 grammar. In C. Chapelle, & S. Sauro (Eds.), *The handbook of technology and second language teaching and learning* (pp. 26–44). John Wiley & Sons.

Isenberg, N. A. (2010). *A comparative study of developmental outcomes in Web-based and classroom-based German language education at the post-secondary level: Vocabulary, grammar, language processing, and oral proficiency development* [Unpublished doctoral dissertation]. The Pennsylvania State University.

Krashen, S. (2014). Does Duolingo "trump" university-level language learning? *The International Journal of Foreign Language Teaching*, *9*(1), 13–15.

Lee, J. S. (2020). The role of grit and classroom enjoyment in EFL learners' willingness to communicate. *Journal of Multilingual and Multicultural Development*, *10*, 1–17. https://doi.org/10.1080/01434632.2020.1746319

Lin, C.-H., & Warschauer, M. (2015). Online foreign language education: What are the proficiency outcomes? *The Modern Language Journal*, *99*(2), 394–397.

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*(3), 293–311.

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. https://doi.org/10.1111/flan.12454

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

Lord, G. (2016). Rosetta Stone for language learning: An exploratory study. *The IALLT Journal*, *46*, 1–35.

Lys, F. (2013). The development of advanced learner oral proficiency using iPads. *Language Learning & Technology*, *17*, 94–116.

Munday, P. (2016). The case for using Duolingo as part of the language classroom experience. *RIED: Revista Iberoamericana de Educación a Distancia*, *19*(1), 83–101.

Munday, P. (2017). Duolingo. Gamified learning through translation. *Journal of Spanish Language Teaching*, *4*(2), 194–198.

Plonsky, L. (2017). Quantitative research methods. In S. Loewen, & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 505–521). Routledge.

Plonsky, L., & Oswald, F. L. (2014). How big is 'big'? Interpreting effect sizes in L2 research. *Language learning*, *64*, 878–912.

Plonsky, L., & Ziegler, N. (2016). The CALL-SLA interface: Insights from a second-order synthesis. *Language Learning & Technology*, *20*(2), 17–37.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Rachels, J. R., & Rockinson-Szapkiw, A. J. (2018). The effects of a mobile gamification app on elementary students' Spanish achievement and self-efficacy. *Computer Assisted Language Learning*, *31*(1–2), 72–89.

Rogers, J. (2017). The spacing effect and its relevance to second language acquisition. *Applied Linguistics*, *38*(6), 906–911.

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

Shintani, N., Li, S., & Ellis, R. (2013). Comprehension-based versus production-based grammar instruction: A meta-analysis of comparative studies. *Language Learning*, *63*, 296–329.

Smith, B. (2017). Technology-enhanced SLA research. In C. A. Chapelle, & S. Sauro (Eds.), *The handbook of technology and second language teaching and learning* (pp. 444–458). Wiley Blackwell.

Soneson, D., & Tarone, E. E. (2019). Picking up the PACE: Proficiency assessment for curricular enhancement. In P. M. Winke, & S. M. Gass (Eds.), *Foreign language proficiency in higher education* (pp. 45–70). Springer. ProQuest Ebook Central. http://ebookcentral.proquest.com/lib/wvu/detail.action?docID=5622543

Spoden, C., Fleischer, J., & Leuchat, M. (2020). Converging development of English as Foreign Language listening and reading comprehension skills in German upper secondary schools. *Frontiers in Psychology*, *11*, 1116. https://doi.org/10.3389/fpsyg.2020.01116

Strawbridge, T., Soneson, D., & Griffith, C. (2019). Lasting effects of pre-university language exposure on undergraduate proficiency. *Foreign Language Annals*, *52*, 776–797.

Sun, Y.-C. (2012). Examining the effectiveness of extensive speaking practice via voice blogs in a foreign language learning context. *CALICO Journal*, *29*, 494–506.

Tarone, E. (2015). Perspectives: Online foreign language education: what are the proficiency outcomes? *The Modern Language Journal*, *99*(2), 392–415.

Tschirner, E. (2016). Listening and reading proficiency levels of college students. *Foreign Language Annals*, *49*, 201–223.

Ushida, E. (2005). The role of students' attitudes and motivation in second language learning in online language courses. *CALICO Journal*, *23*, 49–78.

van Deusen-Scholl, N. (2015). Assessing outcomes in online foreign language education: What are key measures for success? *The Modern Language Journal*, *99*(2), 398–400.

Vandergrift, L., & Baker, S. (2015). Learner variables in second language listening comprehension: An exploratory path analysis. *Language learning*, *65*, 390–416.

Vesselinov, R. (2009). Measuring the effectiveness of Rosetta Stone [White paper]. Rosetta Stone. http://resources.rosettastone.com/CDN/us/pdfs/Measuring_the_Effectiveness_RS-5.pdf

Vesselinov, R., & Grego, J. (2012). The Duolingo efficacy study [White paper]. Duolingo. https://static.duolingo.com/s3/DuolingoReport_Final.pdf24

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

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

Vesselinov, R., & Grego, J. (2018). The italki efficacy study [White paper]. Italki. https://www.gettingsmart.com/wp-content/uploads/2018/01/italki2018FinalReport.pdf

Volle, L. (2005). Analyzing oral skills in voice e-mail and online interviews. *Language Learning & Technology*, *9*, 146–163.

Wallace, M. P. (2020). Individual differences in second language listening: Examining the role of knowledge, metacognitive awareness, memory, and attention. *Language learning*, *49*, 1. https://doi.org/10.1111/lang.12424

Winke, P. M., Zhang, X., Rubio, F., Gass, S. M., Sonenson, D., & Hacking, J. F. (2020). The proficiency profiles of language students: Implications for programs. *Second Language Research & Practice*, *1*(1), 25–64. http://hdl.handle.net/10125/69840

Winke, P. M., Gass, S. M., Soneson, D., Rubio, F., & Hacking, J. F. (2014–2017). Foreign language proficiency test data from three American universities. Inter-University Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/ICPSR37499.v1

Wolf, M. C., Muijselaar, M. M. C., Boonstra, A. M., & de Bree, E. H. (2019). The relationship between reading and listening comprehension: Shared and modality-specific components. *Reading and Writing*, *32*, 1747–1767.

How to cite this article: Jiang, X., Rollinson, J., Plonsky, L., Gustafson, E., & Pajak, B. (2022). Evaluating the reading and listening outcomes of beginning-level Duolingo courses. *Foreign Language Annals*, 1–29. https://doi.org/10.1111/flan.12600

APPENDIX A

Table A1

TABLE A1 Background information of the participants in the Spanish and French courses

Categories	Spanish ($N = 135$)	French ($N = 90$)
Age		
Mean (SD)	46.89 (18.52)	39.54 (14.96)
Prefer not to answer	8	7
Language before age 6		
English	96	70
Early bilingual[a]	13	6
Other[b]	26	14
Highest level of education		
High school	5	7
Associate degree	5	5
Bachelor's degree	51	36
Master's degree	55	28
Doctoral degree	18	14
Prefer not to answer	1	0
Ethnicity		
African American	4	2
Asian	18	9
Caucasian	105	71
Other	4	5
Prefer not to answer	4	3
Gender		
Female	76	36
Male	59	48
Other	0	4
Prefer not to answer	0	2

[a]The early bilinguals learning Spanish ($n = 13$) spoke English and one of the following languages: Chinese (4), Italian (2), Arabic, Hindi, Korean, Norwegian, Polish, Taiwanese, and Turkish. The early bilinguals learning French ($n = 6$) spoke English and one of the following languages: Tagalog (2), Arabic, Spanish, Slovak, and Tamil.

[b]Spanish learners who did not speak English before age 6 ($n = 26$) spoke one of the following languages: Chinese (7), German (3), French (2), Japanese (2), Russian (2), Arabic, Dutch, Farsi, Finnish, Greek, Hindi, Tamil, Turkish, Urdu, and Zulu. French learners who did not speak English before age 6 ($n = 14$) spoke one of the following languages: Spanish (4), Russian (3), Chinese (2), Bengali, Filipino, Indonesian, Italian, and Punjabi/Hindi.

APPENDIX B

The Background Survey

1. What language(s) was/were spoken in your home before you were 6 years old?

   _____

2. What other languages do you speak?

   _____

3. Why are you learning Spanish? (Check all that apply)

   For travel For school For job-related purposes

   For fun/leisure For memory/brain acuteness

   For social purposes Other: _____

4. What other languages have you studied?

   _____

5. What is your highest level of education?

   Some high school High school Associate's degree

   Bachelor's degree Master's degree Ph.D.

   Trade school Prefer not to answer

6. What is your age?

   _____

7. What gender do you identify as?

   Male Female Other: _____ Prefer not to answer

8. Please specify your ethnicity.

   Caucasian African American Latino or Hispanic

   Asian Other: _____ Prefer not to answer

9. How much Spanish do you think you knew before studying on Duolingo?

0 1 2 3 4 5 6 7 8 9 10

Nothing Perfect

10. How much Spanish do you think you know now?

0 1 2 3 4 5 6 7 8 9 10

Nothing Perfect

11. In which area or areas do you think Duolingo helped you the most? (Check all that apply)

Vocabulary Grammar Pronunciation Listening

Speaking Reading Writing

12. How much time (in hours) per week did you use Duolingo to learn Spanish?

_____

13. In addition to the Duolingo Spanish lessons, what other Duolingo resources did you use to learn the language? (Check all that apply.)

Duolingo Spanish Stories

Duolingo Spanish Podcasts

Duolingo Tips in Spanish

Nothing else

14. What do you like about learning Spanish on Duolingo?

_____

15. What do you want to see changed on Duolingo?

_____

16. Did you have experience learning Spanish **before** using Duolingo?

Yes No

–(If yes) How did you learn Spanish before using Duolingo? (check all that apply)

Being around Spanish speakers

High school Spanish classes

College Spanish classes

Language apps

Internet-based materials such as podcasts and YouTube

Textbooks and other materials in print

Other: _____

17. Did you take Spanish classes **during** the time you used Duolingo?

Yes No

18. Did you use other programs or apps to learn Spanish **during** the time you used Duolingo?

Yes No

19. To participate in the study, you will need to take online tests. Do you have access to a webcam-enabled, internet-connected computer?

Yes No

20. To sign you up for the tests and get your official certificates, we'll need your name. Please fill in your name below. Thank you!

First name

Last name

APPENDIX C

Calculating time spent learning

For every exercise that a user completes in a Duolingo session, we store the amount of time taken. To calculate time spent learning, we sum the amount of time taken per exercise with a maximum of 60 s per exercise. We apply this maximum because users can close the app during an exercise, and re-open it much later, and we do not want to include the large period of time in which the user was not using the app. Nearly all exercises are completed within 60 s, so this does not have a large impact on the final result.

This approach to measuring learning was fully developed at the beginning of August 2019. Unfortunately, some participants in this study began using Duolingo as early as 2012. Therefore, we had to resort to a less systematic approach where we measured the wall-clock time that a user spent in sessions, subject to a 10-min cutoff per session. To transform this value to be consistent with our normal method for measuring time spent learning, we multiplied the values with a constant that had been previously calculated with a linear regression model. In previous research, we had found that this approach produces highly correlated results with our usual approach.

To verify this, we computed the time spent learning using both methods on all data collected from August 2019 to July 2020 for participants in this study. We found a Pearson correlation of 0.99 and an average difference of 2.9% between the two methods.