Language Learning & Technology
ISSN 1094-3501 CC BY-NC-ND

2024, Volume 28, Issue 1
pp. 1–26

**ARTICLE**

# The effectiveness of Duolingo in developing receptive and productive language knowledge and proficiency

*Bryan Smith*, Arizona State University*
*Xiangying Jiang, Duolingo*
*Ryan Peters, Duolingo*

## Abstract

*This study aimed to evaluate the effectiveness of Duolingo's Spanish course for English speakers over a three-month period for independent learners (n = 48). We examined learning outcomes in general proficiency, reading, writing, listening, speaking, vocabulary, grammar, and pronunciation. We also explored the relationship between learner usage, experience factors, and self-reported experiences using the app. After about 27 hours of study, participants significantly improved in all receptive and productive ability measures, supporting the notion that language learning apps can enhance a range of language skills. Session completion, accuracy rate, and positive user experience were linked to this observed growth.*

***Keywords:*** *Duolingo, Language Learning Apps, Receptive and Productive Skills, Language Proficiency*

***Language(s) Learned in This Study:*** *Spanish*

**APA Citation:** Smith, B., Jiang, X., & Peters, R. (2024). The effectiveness of Duolingo in developing receptive and productive language knowledge and proficiency. *Language Learning & Technology*, *28*(1), 1–26. https://hdl.handle.net/10125/73595

## Introduction

Smartphones and tablets have transformed language learning, providing easy access to educational apps and resources anytime, anywhere. This enables more consistent and potentially meaningful exposure to the target language, which is essential for successful language learning. Many apps also use gamification to make learning engaging and fun, boosting user engagement (Shortt et al., 2023).

Online and mobile-based language learning (MALL) offers a unique learning environment that transcends geographical barriers, making it accessible to learners in nearly any situation (Loewen et al., 2020). This accessibility broadens the instructional contexts of second language acquisition (SLA) to include online courses, mobile-assisted learning (including app-based instruction) and traditional classroom teaching. These developments affect SLA by influencing key topics such as the nature and impact of input, learner interaction and engagement with course content, and the corrective feedback they receive (Collentine & Freed, 2004).

With mobile-based language learning, instruction moves beyond the physical classroom and into a digital space where learners may interact with one another or autonomously interact with course materials and gamified elements. The remarkable expansion of mobile language learning content provided by both educational institutions and commercial entities has garnered significant interest in the effectiveness of such products. Despite the popularity of MALL, many language teaching professionals remain skeptical about the effectiveness of such courseware, as noted by Brown (2023) and Lin and Warschauer (2015). This doubt primarily stems from the limited research-based evidence that supports the effectiveness of

*** Corresponding Author:** Bryan Smith, bryansmith@asu.edu

mobile-based learning through standardized and validated language proficiency measures (Burston & Giannakou, 2022; Tarone, 2015). There is also a perception among many researchers and practitioners that app companies oftentimes make exaggerated claims about their product.

Understanding the effectiveness of popular language apps is vital in MALL research. Duolingo, for instance, provides free and premium courses via web and mobile apps, dominating about 90% of the global user base in the $3 billion app-based language learning market (Curry, 2023; M Science, 2023). This study aims to assess the impact of Duolingo on independent, informal language learning over three months, using standardized proficiency tests and other measures of language ability. The literature review that follows first defines and then explores the evidence for the effectiveness of MALL in general and then turns to findings regarding user experience and language learning in app-based environments.

## Literature Review

### Definition of MALL

MALL is defined as "learning a second or foreign language through the use of one or more of various mobile devices including, but not restricted to, mobile phones (including smartphones), tablets, personal digital assistants (PDAs), MP3/MP4 players, electronic dictionaries, and gaming consoles" (Stockwell, 2022, p. 8). MALL research involves using mobile-based apps or device features for language teaching and learning in specific environments with specific participants and treatment conditions (Burston & Giannakou, 2022). Studies range from examining WeChat, a popular Chinese social media messaging platform, for learner interaction (Li, 2018) to independent learning on apps like Duolingo (Sudina & Plonsky, 2024). MALL can supplement traditional classroom learning or provide for independent, extramural learning. Indeed, the wide variety of MALL learning contexts complicates generalizing or comparing study results.

### Effectiveness of MALL

Since the early 1990s, over 3,500 MALL-related studies, overviews, and meta-analyses have been published (Burston & Giannakou, 2022). Many have shown its benefits, which include improved accessibility, learner motivation, vocabulary acquisition, and personalized learning. A meta-analysis by Burston and Giannakou (2022) found significant MALL impacts in both between-group (mobile vs. non-mobile learners; $k = 84$) and within-group (pre-post designs; $k = 56$) studies, with effect sizes of $g = .72$ and $g = 1.16$, comparable to other SLA areas (Plonsky, 2017).

While evidence supports MALL effectiveness (Chen et al., 2020), researchers have identified some design-related shortcomings, making the findings hard to interpret. These include inadequate establishment or reporting of baseline language proficiency (Burston & Athanasiou, 2020; Burston & Giannakou, 2022) and a primary focus on vocabulary (34%), reading (20%), and grammar (12%) (Burston & Giannakou, 2022), which may reflect a belief that apps like Duolingo are mainly effective for receptive skills. Many also have short treatment durations, with only about half lasting 8 weeks or more, and lack verification of actual app usage during the study.

Another major issue in MALL research design is regarding instrumentation. Most studies use researcher-developed tests to assess instruction quality (Loewen et al., 2020). These tests, often chosen for convenience (Park et al., 2022; Thomas, 2006), may lack robust validity, making conclusions less reliable and possibly obscuring comparison with established proficiency scales like the American Council on the Teaching of Foreign Languages (ACTFL). Nonetheless, an increasing number of studies, particularly in app-based learning, are employing standardized assessments to evaluate MALL's effectiveness (e.g., Jiang et al., 2021; Loewen et al., 2020). A combination of custom-made and standardized tests may be considered the best approach, as both have strengths and weaknesses. This study adopts this approach to leverage the flexibility and specificity of custom-made tests while ensuring comparability and reliability through standardized assessments.

## Effectiveness of Language Learning Apps

Currently, there are no published meta-analyses specifically on app-based MALL, and quality research on these apps' effectiveness is also scarce. However, extensive commissioned studies, primarily by Vesselinov and Grego (e.g., Vesselinov, 2009; Vesselinov & Grego, 2012, 2016a, 2016b, 2018), are available as white papers on company websites. These include evaluations of Rosetta Stone, Duolingo, Babbel, Busuu, and Italki focusing on their Spanish courses' effectiveness. These studies, using a pretest-posttest design and the Web-based Computer Adaptive Placement Exam (WebCAPE), assess vocabulary, reading, and grammar. They measure effectiveness through score improvements and study hours, providing estimates to match a first-semester university language course level. Nonetheless, these results lack a broader context and rigorous methodology, such as initial proficiency control and time on task, and could benefit from more comprehensive proficiency assessments.

Recent studies have examined the effectiveness of Babbel and Duolingo (Kessler et al., 2023; Loewen et al., 2019, 2020). Loewen et al. (2020) found Babbel improved oral proficiency, grammar, and vocabulary among 54 university students, linking learning gains to usage time and interest in Spanish. A study on beginner Turkish learners using Duolingo (Loewen et al., 2019) reported better written than oral skill development. Kessler et al. (2023) compared adult learners using Babbel or Duolingo for Turkish over eight weeks, noting similar improvements in various language skills, but no significant differences between the apps.

Studies comparing online language platforms such as Rosetta Stone and Duolingo to traditional classroom instruction reveal no clear drawbacks for online methods. Lord (2015, 2016) found comparable Spanish language skills development between university beginners using either Rosetta Stone or traditional teaching over 16 weeks, despite traditional classes demanding more study time. Rachels and Rockinson-Szapkiw (2017) also saw no significant disparity in vocabulary and grammar progress between elementary students learning Spanish through Duolingo and those in conventional classes, highlighting Duolingo's effectiveness at this educational level.

Jiang et al. (2021) assessed Duolingo users completing Basic Spanish and French courses, targeting the Common European Framework of Reference (CEFR) A2 level. Focusing on individuals with minimal prior language knowledge and using solely Duolingo, the study found learners reached Intermediate Low in reading and Novice High in listening as per ACTFL tests. These achievements paralleled those of university students after four semesters (Winke et al., 2020), yet Duolingo users needed only about half the time. However, the study focused only on receptive skills without a pretest for initial proficiency, limiting precision in learning gain measurements. A pretest could have provided a more accurate evaluation.

In sum, language learning apps can match traditional methods in linguistic improvement. Yet, many studies have not assessed overall proficiency using standardized tests and almost exclusively focused on receptive skills. Addressing these gaps, this study uses a standardized proficiency test along with other tools to evaluate a broader spectrum of language skills, both receptive and productive.

## User Experiences in App-based Learning

Research on language learning apps includes user perceptions, attitudes, and motivation. User experience, involving reactions, ease of use, content relevance, and engagement, shows mixed outcomes (He & Loewen, 2022; Kessler, 2021; Kohnke, 2020; Loewen et al., 2019, 2020; Rosell-Aguilar, 2018; Sporn et al., 2020). Babbel users valued its content but faced declining motivation (Loewen et al., 2020), Duolingo users liked gamification despite repetitive tasks (Byrne et al., 2022; Kessler, 2021; Loewen et al., 2019), and Busuu users sometimes felt irritated (Rosell-Aguilar, 2018). A common trend is a gradual drop in initial enthusiasm for app usage (Peng et al., 2021).

User feedback is vital for improving apps but should be paired with learning outcome assessments.

However, few studies link user experience with learning effectiveness, despite the evidence that nonlinguistic factors like motivation and self-confidence positively correlate with learning outcomes (Clément et al., 1994; Gardener, 1985; Masgoret & Gardner, 2003; Yu et al., 2023). Thus, exploring user experience in relation to app-based learning effectiveness, especially in self-directed contexts, is crucial. Factors such as motivation and perceived efficacy are key for sustained engagement, necessary for language development (Loderer et al., 2020).

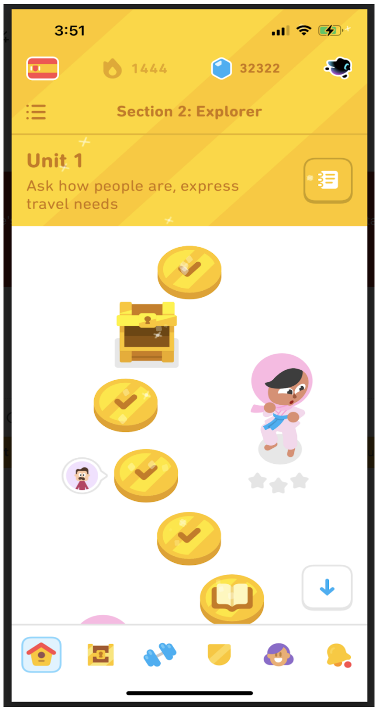
## The Present Study

This study aims to explore linguistic improvements of non-Spanish speaking university students who used Duolingo for Spanish learning over three months, examining user experience during this period. It addresses previously identified design issues in app-based language learning research. A third-party standardized test was employed to assess gains in receptive and productive skills and overall proficiency. Additional measurements of pronunciation, grammar, and vocabulary were conducted using researcher-developed tests, adapted from the literature. The study established baseline proficiency levels, controlled for time on task and learner usage patterns across twelve weeks, and separated language app usage from formal instruction.

Participants engaged with the Duolingo Spanish curriculum for English speakers, aligned with the CEFR, a widely recognized language proficiency standard (Council of Europe, 2001). The course is structured into sections, each with varying numbers of units. For example, Section 1 has 5 units, while Section 2 contains 15. Units include lessons introducing new content or reviewing previous material, along with Stories for reading and listening practice. Figure 1's screenshot shows the beginning of Unit 1 in Section 2, where each circle represents four lessons and a final review. An open book icon indicates a Story. This unit has seven circles (28 lessons, 7 reviews) and two Stories.

**Figure 1**

*Screenshot of Duolingo Course Structure*



*Note*. Reprinted with permission from Duolingo

Activities within the lessons target various language skills including vocabulary, grammar, reading, listening, writing, and speaking. The platform enhances listening and speaking development by offering a substantial amount of target language input and opportunities for language production. All course materials are supported by audio, allowing learners to adjust the playback speed, a feature that can influence the difficulty level of listening exercises (East & King, 2012). Additionally, speech recognition technology is used in speaking exercises to provide feedback. The inclusion of Stories facilitates reading and listening comprehension practice at a discourse level, helping to contextualize lesson content in everyday scenarios and offering extended opportunities for skill development.

## Research Questions

Utilizing a within-subject quasi-experimental pretest-posttest research approach, the study posed the following research questions (RQs):

RQ1: To what extent can English-speaking university students improve their overall Spanish language proficiency, skill area proficiency, and other linguistic abilities by using Duolingo Spanish independent of any formal instruction for three months?

RQ2: What is the relationship (if any) between learner usage factors (e.g., time spent learning, number of sessions completed, and session accuracy rate), user experience factors (e.g., interest, perceived efficacy, and motivation), and gains in overall Spanish language proficiency, skill area proficiency, and other linguistic abilities?

RQ3: What were the other reported experiences of these L2 Spanish learners while using Duolingo?

## Methods

### Participants

This study recruited participants from a large public university in the Southwest United States, following approval by the university's Institutional Review Board (IRB). An invitation was disseminated via departmental listservs to active students, from which 245 expressed interest by completing an eligibility survey. The initial exclusion criteria were: speaking Spanish at home before the age of six, current enrollment in a Spanish course, completion of advanced Spanish classes (200-level or above), or pursuing a major/minor in Spanish. Seventy-four students meeting these criteria were invited for a pretest. Further exclusion at the pretest stage involved scoring above an overall proficiency level of 4 on the **STA**ndards-based **M**easurement of **P**roficiency (STAMP) 4S Spanish Test by Avant Assessment (detailed in the Instruments section). The study required participants to have a STAMP Level of 4 (Intermediate Low) or below in order to align with Duolingo's content coverage up to the intermediate level.

Of the 69 participants who commenced the learning process, 48 completed the study, meeting the set engagement criterion of at least 15 minutes of daily app usage for a minimum of five days per week, a commitment agreed upon in their consent forms. The final participant demographic included: 30 monolingual English speakers, 7 bilinguals (English plus another language), and 11 non-native speakers of English. Based on the pretest, Spanish proficiency levels were: Novice Low (14), Novice Mid (14), Novice High (18), and Intermediate Low (2), with a general trend of stronger reading and listening skills over speaking and writing skills (detailed in Table 2).

Participants joined the study in three cohorts starting on January 28, February 25, and March 10, respectively. The first cohort, consisting of 21 on-campus students, was tested in a computer lab, while the second (17 online students) and third cohorts (10 on-campus students) completed their tests remotely with live proctoring. Despite these logistical variations, analysis revealed no significant differences in pretest scores or Duolingo usage among these groups.

## Instruments

### *Eligibility Questionnaire*

The initial eligibility survey, with 245 respondents, gathered information on their linguistic background, self-rated Spanish skills, learning experiences, motivation, and opinions on app-based language learning. This information was used to determine who would be invited to the next round of the study (the pre-test).

### *Proficiency Test: STAMP 4S Spanish Test*

Pre-qualified participants took the STAMP 4S Spanish Test[1] by Avant Assessment to establish baseline proficiency. This online, ACTFL-aligned computer-adaptive test evaluates reading, writing, listening, and speaking skills, on the following scales: writing and speaking (0-8), reading and listening (0-9), and overall proficiency (0-9), corresponding to the nine ACTFL sublevels from Novice to Advanced (Table 1). After three months on Duolingo, participants retook this test as a post-test assessment.[2]

**Table 1**

*Alignment of the STAMP Scale with ACTFL Proficiency Scale*

| Stamp Level | ACTFL Proficiency Level | Sublevel |
|---|---|---|
| 1 | Novice | Low |
| 2 | Novice | Mid |
| 3 | Novice | High |
| 4 | Intermediate | Low |
| 5 | Intermediate | Mid |
| 6 | Intermediate | High |
| 7 | Advanced | Low |
| 8 | Advanced | Mid |
| 9 | Advanced | High |

*Note*. Adapted from Santos (2022, p. 2)

Individual STAMP level ratings were assigned for reading, listening, speaking, and writing skills, as well as overall proficiency, which was calculated as the average of the four skill ratings. To simplify the discussion of gain scores, we refer to the STAMP measures as "Proficiency" measures moving forward.

### *Linguistic Measures*

In addition to the STAMP 4S test, we adapted three other measures of language knowledge (hereafter, linguistic measures), consisting of a vocabulary test, a grammar test that included error identification and correction, and an elicited imitation task (EIT). Following Sudina and Plonsky (2024), who noted that "full range" tests may not be sensitive enough to adequately capture learning gains of lower-level learners, we revised these tests to only include language sampled from Duolingo's CEFR-aligned curriculum up to the first half of the B1 level. Each test is described in the following section.

### Vocabulary Test

The vocabulary test, adapted from LEXTALE-ESP by Izura et al. (2014), consisted of 75 Spanish-looking letter sequences, including 50 real words and 25 nonwords. The words were ordered from easy to difficult, with nonwords randomly inserted. Participants were asked to indicate if the sequences were Spanish words. The test was revised to feature intermediate or lower-level words, shortened from 90 to 75 items, and converted to an online format via Qualtrics. The revised test contained 35 of the items from the original test. Following Izura et al. (2014), learners scored +1 point for each correct real word identified and -2 points for each incorrect nonword identification with a score range of -50 to 50. The test, used for both pretest and posttest, was automatically scored. The internal consistency reliability coefficients (KR-21) were 0.93 on pretest and 0.90 on posttest (See Izura et al., 2014 for more information on validation of the LEXTALE-ESP). The word list can be found here in the Instruments for Research into Second Languages (IRIS) database.

### Grammar: Error Identification and Correction

For the grammar test, we adapted the instrument developed by Leonard and Shea (2017). The test was revised by the researchers to include features only at the intermediate or lower-level. This resulted in a 30-item test that retained 11 sentences from the original, each containing one grammatical error (verb tense, aspect or mood, adjective–noun agreement, or incorrect pronoun). Administered via Qualtrics, participants were asked to identify the error by highlighting it and then correct it by typing in the correction, earning one point each for accurate identification and correction. Responses were scored automatically and reviewed for accuracy by one of the researchers. The same test was used for both pretest and posttest, with scores ranging from 0 to 30 for identification and correction, for a total grammar score between 0 and 60. KR-21 internal consistency reliability coefficients were as follows: error identification (0.80 pretest, 0.73 posttest), error correction (0.71 pretest, 0.68 posttest), and total (0.88 pretest, 0.89 posttest) (See Leonard & Shea, 2017 for more information on the validation of the test). The items can be found here in the IRIS database.

### The Elicited Imitation Task (EIT)

The EIT Task was adapted from Solon et al. (2019) and shortened from 36 sentences to 25 and revised for intermediate or lower-level learners. In this task, participants listened to and then repeated each sentence. The revised version contained 10 sentences that were borrowed word-for-word from the original, with another five sentences only very slightly modified for vocabulary level. Thus, about 60% of the EIT came directly from Solon et al. (2019). The same test was used for both pretest and posttest.

Audio input alternated between male and female native Spanish speakers. A short pause and tone followed each sentence, signaling participants to repeat and audio-record their responses. Participants clicked on a “record” button to begin their recording and a “stop” button to end their recording. There was no time limit for responses. The task was administered via Qualtrics with an audio recording function provided by phonic.ai. Two Spanish native speaker raters were trained on the scoring rubric selected from Solon et al. (2019) and engaged in an anchoring session with one of the researchers. One rater was a PhD student at the host university and had over ten years Spanish teaching experience. The other rater was a very experienced Spanish teacher, working for Duolingo as a freelancer. Outputs were rated on a 0-4 scale, with a maximum possible score of 100 (the item list and rating criteria can be found here in the IRIS database.

After an initial rater anchoring session, the two raters independently scored a set of 250 audio recordings. The inter-rater reliability between the raters’ scores on these recordings was a Pearson *r* of 0.91. We identified seven recordings with a rating difference greater than one point, and after raters’ independent review of their scores of these items, the reliability increased to 0.94. One rater then scored all recordings, unaware of their pretest or posttest status. The internal consistency reliability coefficients (KR-21) of the test were 0.93 and 0.90 for pretest and posttest, respectively (See Solon et al., 2019 for more information on the validation of the EIT test).

### *Posttest Questionnaire*

The posttest questionnaire asked questions about participants' self-perceived Spanish proficiency, their likelihood to continue studying Spanish or other languages on Duolingo, and their self-confidence and efficacy. Statements were rated on a 1-5 Likert scale (strongly disagree to strongly agree), while likelihood questions used a 0-5 scale (not likely at all to extremely likely). Participants' ratings on these items helped identify user experience factors and their relation to language development. Open-ended questions gathered insights on participants' Duolingo experiences. The questionnaire was developed by the researchers for this study and can be found here in the IRIS database.

### *App Usage*

To explore the connection between app usage and language development, learners' app data were collected using the Duolingo for Schools application. These data included total minutes spent on Duolingo during 12 weeks, average minutes per week, numbers of lessons, level reviews, Stories completed during the 12 weeks, and the average accuracy rate per session.

### Data Collection Procedures

Pre-qualified participants took a pretest battery, including the proficiency and linguistic measures described above. Two hours were allotted for the pretest. On average participants took about 90 minutes to complete all sections. Again, those scoring above STAMP Level 4 on overall proficiency (Intermediate Low) were disqualified.

Participants joined Duolingo for Schools, a free platform for educators. It offered online classrooms and allowed for progress monitoring and data collection. Participants started at the beginning of the Spanish course, with the option to advance. They were asked to study at their own pace for 12 weeks, at least 15 minutes daily, five days a week, receiving weekly reminders. Posttests were in Week 13. On average, it took participants about 90 minutes to complete all sections of the posttest, including the posttest questionnaire. The first 21 participants completed proctored tests in a university lab; others did so online with remote proctors for the STAMP 4S Test. Participants who completed all aspects of the study received a $200 electronic gift card.

## Results

### Research Question 1: Gains on proficiency and linguistic knowledge, both receptive and productive

We used a series of Wilcoxon signed-rank tests to compare pre- and post-test mean scores in the STAMP 4S Spanish Test. These non-parametric tests were used due to the ordinal nature of the scores (1-9) and that the pre- post-test measures were on paired groups. The assumptions of this test were met. Table 2 displays the descriptive statistics and test results.

In the pretest, two participants had NR (not ratable)[3] speaking scores, and in the posttest, one had an NR writing score, while three others had NR speaking scores. These individuals were excluded from the analysis. Wilcoxon tests revealed significant improvements in overall Spanish proficiency after three months of Duolingo use, with a large effect size (*r*=0.61) and progression from Novice-Mid to Novice-High. All four skills showed significant improvements with large effect sizes. Appendix A visually presents these changes.

**Table 2**

*Descriptive Statistics and Wilcoxon Signed-rank Test Results on Proficiency Measures*

| | | Pretest | | *Posttest* | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Ability tested* | *N* | *M* | *SD* | *M* | *SD* | *z* | *p* | *Wilcoxon Signed Rank Test Prob.* | *r* |
| Overall proficiency | 47 | 2.60 | 0.85 | 3.74 | 0.89 | -5.93 | <.00001 | 4.0 | .61 |
| Reading | 37 | 4.29 | 1.29 | 5.71 | 1.15 | -5.14 | <.00001 | 11.0 | .60 |
| Listening | 28 | 3.54 | 0.97 | 4.19 | 0.94 | -4.12 | <.00001 | 22.2 | .55 |
| Speaking | 28 | 1.29 | 1.18 | 2.53 | 1.31 | -4.62 | <.00001 | 0.00 | .62 |
| Writing | 35 | 1.34 | 1.26 | 2.74 | 1.37 | -4.61 | <.00001 | 33.5 | .55 |

*Note.* Following Plonsky and Oswald (2014), values for "*r*" in the L2 research using within-groups designs should generally be interpreted as .25 (small), .40 (medium), and .60 (large).

To address gains on the other measures of linguistics knowledge (vocabulary, grammar-error-identification, grammar-error-correction, grammar-total, and EIT speaking), a series of paired t-tests were performed on pre- and post-test mean scores. Receptive skills were evaluated through vocabulary and grammar-error-identification, while productive skills were assessed using grammar-error-correction and EIT speaking. Histograms for all linguistic pre- and post-test measures were examined by the researchers and the distribution was determined to be normal. Table 3 displays the descriptive statistics and test results. All five linguistic measures tested showed significant improvements with effect sizes (Cohen's *d*) ranging from 0.69 - 1.11 (small to medium). Appendix A visually presents these changes.

**Table 3**

*Descriptive Statistics and Paired t-Test Results on Linguistic Measures (N = 48)*

| | *Pretest* | | *Posttest* | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Ability tested* | *M* | *SD* | *M* | *SD* | *MD* | *95% CI* | *t (df)* | *p* | *d* |
| Vocabulary | 10.58 | 9.36 | 17.73 | 9.73 | -7.15 | [-10.15, -4.14] | -4.79(47) | <.001 | 0.69 |
| Grammar error-identification | 7.10 | 5.28 | 11.21 | 5.50 | -4.10 | [-5.79, -2.42] | -4.90 (47) | <.001 | 0.71 |
| Grammar error-correction | 3.58 | 3.31 | 7.94 | 4.42 | -4.35 | [-5.79, -3.22] | -7.72 (47) | <.001 | 1.11 |
| Grammar total | 10.69 | 7.80 | 19.15 | 9.45 | -8.46 | [-10.87, -6.05] | -7.07 (47) | <.001 | 1.02 |
| EIT speaking | 22.77 | 12.09 | 35.39 | 14.41 | -12.52 | [-16.21, -8.83] | -6.83 (47) | <.001 | 0.97 |

*Note.* Following Plonsky and Oswald (2014), values for "*d*" in the L2 research using within-groups designs should generally be interpreted as .60 (small), 1.00 (medium), and 1.40 (large).

## Research Question 2: The relationship between learner usage factors, user experience factors, and learning outcomes

### *Learner Usage Factors and Learning Outcomes*

Usage data was obtained from the Duolingo for Schools platform, including total learning time, weekly learning time, total number of sessions, and average session accuracy[4] rate. Table 4 presents the means and standard deviations for these factors.

The analysis to answer Research Question 2 consisted of two steps. The first step investigated the association between usage factors and proficiency measure gains. The second analysis focused on these same usage factors and the other linguistic measure gains.

**Table 4**

*Descriptive Statistics on Usage Factors*

| Usage factors | M | SD |
|---|---|---|
| Total time learning (minutes) | 1608.58 | 676.52 |
| Weekly learning time (minutes) | 134.13 | 56.33 |
| Number of lessons | 231.25 | 82.26 |
| Number of level reviews | 58.13 | 19.92 |
| Number of Stories | 44.54 | 17.91 |
| Total number of sessions | 333.92 | 110.80 |
| Session accuracy rate (%) | 93.00 | 3.11 |

On average, participants spent 26.81 hours learning on Duolingo over three months, averaging 2.23 hours weekly. In subsequent analyses, only the weekly learning time was used, as it was derived from the total learning time. Total number of sessions was used in the analysis because it is the sum of the number of lessons, level reviews, and Stories.

Linear regression models were employed to explore the impact of usage factors on post-test scores for linguistic and proficiency measures, while controlling for corresponding pre-test scores. Table 5 details all significant predictive relationships.

**Table 5**

*Usage Factors Predicting Learning Outcomes*

| Measure | $R^2$ | Estimate | *SE* | *p* | 95% CI | |
|---|---|---|---|---|---|---|
| | | | | | *LL* | *UL* |
| Linguistic | | | | | | |
| Vocabulary | .298 | | | | | |
| Sessions | | .047 | .018 | .011 | .011 | .083 |
| Grammar ID | .403 | | | | | |
| Sessions | | .020 | .009 | .029 | .002 | .037 |
| Accuracy | | .454 | .212 | .038 | .027 | .882 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Grammar Total | .458 | | | | | |
| Sessions | | .030 | .015 | .045 | .001 | .059 |
| Proficiency | | | | | | |
| Reading | .397 | | | | | |
| Accuracy | | .113 | .049 | .026 | .014 | .211 |
| Speaking | .419 | | | | | |
| Accuracy | | .132 | .059 | .030 | .014 | .251 |

Linear regression models revealed that the total number of sessions completed significantly predicted vocabulary, grammar error identification, and overall grammar performance. Session accuracy rate significantly predicted reading, and speaking, as well as grammar error identification. However, weekly learning time was not associated with performance.

### *User Experience Factors and Learning Outcomes*

To examine the relationship between user experience and linguistic outcomes, an exploratory factor analysis (EFA) was conducted based on participants' posttest survey responses, which consisted of 22 questions answered on a five-point Likert scale.[5] In assessing the reliability of our post-treatment questionnaire, we computed Cronbach's Alpha and obtained a value of 0.96 (corrected total item correlations, 0.24-0.86), indicating high internal consistency among the questionnaire items. The assumptions pertinent to factor analysis, including the adequacy of the sample size (Kaiser-Meyer-Olkin Measure of Sampling = 0.855), the sphericity of the dataset (Bartlett's Test of Sphericity = Approx. Chi-Square, 976.78, df = 231, p < 0.001), and the normality of variables (visually inspected), were considered. We recognize the limitations posed by the modest sample size; however, we felt that an EFA may provide valuable data for future research.

Three factors emerged as detailed in Appendix B.[6] The proportion of variance for each factor can be seen in Table 6. The first factor, *Perceived Efficacy*, represents the perceived effectiveness of Duolingo in teaching and improving language skills, as well as the sense of personal achievement and the practical application of the learned language. The second factor, *Continued Engagement*, encompasses the likelihood of continuing to learn Spanish, specifically with Duolingo. The third factor, *Positive User Experience*, captures enjoyment of the learning process, motivation, the likelihood of recommending Duolingo to others, and a sense of pride in using the platform.

**Table 6**

*Proportion of Variance for each Factor*

| | F1 Perceived Efficacy | F2 Continued Engagement | F3 Positive User Experience |
|---|---|---|---|
| SS loadings | 6.24 | 4.74 | 3.68 |
| Proportional Var | 0.28 | 0.22 | 0.17 |
| Cumulative Var | 0.28 | 0.50 | 0.67 |

After identifying the factors, factor-based composite variables were created (see Table 7 for average ratings and standard deviations). These variables were used in mixed-effects models to predict posttest scores on linguistic and proficiency measures, controlling for corresponding pretest scores. Mixed-effects modeling revealed that Positive User Experience significantly predicted EIT speaking (p = .001, Coef.

7.335, CI [2.825, 11.845] and overall proficiency (p = .034, Coef. 0.347, CI [0.026, 0.669]. No other significant relationships were found.

**Table 7**

*Means and Standard Deviations on Factor-based Composite User Experience Variables*

| User experience variable | M | SD |
|---|---|---|
| Perceived efficacy | 3.48 | .18 |
| Continued engagement | 3.74 | .25 |
| Positive user experience | 3.58 | .34 |

*Note.* Mean scores reflect items with factor loadings of 0.50 or greater (see Appendix B)

## Research Question 3:  User experiences with learning Spanish with Duolingo

The final research question examined participants' reflections on their 3-month Duolingo Spanish learning experience. Appendix C reports average ratings and standard deviations for each item. In general, participants found Duolingo enjoyable and effective for learning Spanish, and their motivation to continue increased after three months.

The survey also featured seven open-ended questions. We performed a thematic analysis using a ChatGPT-4 chatbot (OpenAI, 2023), inputting the questions and responses. We asked the chatbot to generate themes based on participant responses (mentions) to each question. We also prompted the chatbot to identify which responses were used to generate each theme. One of the researchers then reviewed the generated themes for accuracy and the themes were determined to be reasonable based on the student responses. Using the list of themes generated for each question, a research assistant then independently coded all student responses for these questions. The simple percentage agreement between ChatGPT-4 and the human rater ranged from 79% (Question 9) to 98% (Question 14). The four main themes generated by ChatGPT-4 for each open-ended question and number of mentions are found in Appendix D.

Taken together, responses suggest that students measure the effectiveness of a language learning app by their ability to apply newly gained knowledge to real-life situations across a broad spectrum of language skills and abilities. They value features that enable meaningful conversations, present challenges, offer feedback to refine accuracy, and support the development of comprehensive language skills through consistent daily practice. Students appreciated Duolingo for its individualized pace, which allowed for personal progress without the constraints of a traditional classroom setting. They highlighted the platform's role in enhancing vocabulary and basic grammar understanding, though they noted a desire for more practical listening skills and pronunciation and more attention to language varieties and cultural nuances. Learners appreciated the gamified learning environment and user-friendly interface, citing the interactive and engaging exercises, such as listening prompts, Stories, and varied activities across all skill areas. The app's accessibility, allowing for on-demand learning alongside the motivational aspects such as progress tracking, was particularly valued.

However, students also noted areas where they would like to see further development. These areas included more grammar instruction, more interactive and challenging content, including real-life conversational practice, and a reduction in the repetitiveness of content. Comments also included requests for better pronunciation guidance, a broader variety of content, and more customizable learning paths, which reflects a demand for a more comprehensive learning experience. Despite these noted areas for improvement, the overwhelmingly positive feedback from those who completed the study highlights the potential effectiveness of app-based language learning environments in creating accessible, engaging, and effective learning experiences that resonate with university-level foreign language learners.

## Discussion

This study assessed Duolingo's impact on learners' Spanish language proficiency and linguistic knowledge over three months. Utilizing a pre- and post-test design, it measured gains in overall proficiency and specific skills such as reading, speaking, writing, and listening, showing significant progress and large effect sizes. It also evaluated vocabulary, grammar, and EIT speaking tests, noting considerable improvements with small to medium effects, aligning with effect sizes from Burston and Giannakou’s (2022) meta-analysis on MALL interventions.

After about 27 hours of study (approximately 2.25 hours weekly), participants significantly improved both receptive (reading and listening) and productive (speaking and writing) skills. Gains in productive skills matched or surpassed those in receptive skills, indicating language apps like Duolingo can enhance productive as well as receptive abilities, grammar, and vocabulary. This might be partly because learners had lower pretest scores in speaking and writing, providing more room for improvement.

Most previous research on language apps has not focused on overall proficiency or productive skills development. For example, Jiang and colleagues only evaluated the reading and listening proficiencies of Duolingo learners in its Spanish and French courses (Jiang et al., 2021) and English courses (Jiang, Peters, et al., 2024). One exception is Loewen et al. (2020), who studied Spanish learners using Babbel. In their three-month study, users logged 11.61 hours on average and improved by 0.7 of an ACTFL level in speaking. In contrast, our study found that after 26.8 hours on Duolingo, participants gained more than one sublevel in overall proficiency, reading, speaking, and writing, with listening skills improving by 0.65 sublevel, advancing from Novice High to Intermediate Low. Thus, our results are consistent with Loewen et al.’s (2020) findings for speaking and provide encouraging insights, which researchers should attempt to replicate.

The second research question examined how usage and user experience factors predict proficiency and language skills improvements. Analysis showed that more sessions predicted better vocabulary, grammar error identification, and overall grammar scores. Session accuracy rate was a strong predictor for reading, speaking, and grammar error identification, emphasizing its importance as a metric for gauging a learner's grasp of the course material. However, weekly learning time didn't significantly relate to outcomes. Contrary to Loewen et al. (2020), where Babbel study time strongly predicted outcomes, Duolingo showed a weak time-outcome correlation, a finding consistent with Sudina and Plonsky (2024). The weak correlation between time spent on Duolingo and outcomes could be due to learners engaging in unproductive activities, like repeating sessions for easy Experience Points (XPs). Second, it might have to do with the “minimum threshold of app use” imposed by the researchers in this study. That is, all learners had to complete a minimum of 15 minutes of active app use per day for at least five days per week, which resulted in a similar amount of app usage across each participant. This creates a floor effect that may lessen the chance of finding an effect for learning time. Additionally, at the session level, the more mistakes learners make, the longer the session will last because learners have to correct their mistakes until they get all the items correct.

Survey analysis highlighted three user experience aspects: Perceived Efficacy, Continued Engagement, and Positive User Experience, with the latter being crucial in self-directed learning. Learners enjoying a better user experience on Duolingo achieved greater proficiency and speaking ability, reinforcing the value of a positive user experience as suggested by prior studies (Clément et al., 1994; Loderer et al., 2020). This study did not show significant associations between learning outcomes and perceived efficacy or continued engagement, perhaps because the commitment of these learners, though surely motivated by the anticipated payment in many cases, was perceived to be generally high. Nevertheless, the role of these variables should be further explored in future studies.

The third research question revealed that users found Duolingo enjoyable and effective for learning

Spanish, retaining a high level of motivation after three months. Participants appreciated Duolingo's gamified, user-friendly, and flexible approach, which helped them improve their Spanish skills through consistent daily practice. However, they also noted the lack of practical listening, speaking, pronunciation, and conversational support, customization, and cultural insights. Learner observations regarding speaking and pronunciation are curious as some of the strongest gains were made in these very areas. Language app developers should take note of these perceived shortcomings.

## Limitations and Future Directions

There are a few limitations of this study which should be addressed in future research. The first is the relatively small sample size, as well as potentially undetected differences between the cohorts of participants. Also, we did not analyze differences in performance between students based on their language background (monolingual English speakers, bilinguals, and non-native speakers of English) or Spanish proficiency level (Novice Low, Novice Mid, Novice High, and Intermediate Low).

Another limitation was the high attrition rate of participants. During a 3-month learning period, 30% of participants left or were removed from the study, which suggests that the remaining participants were highly engaged. Also, the findings based on Spanish learners might not be generalizable to learners in other Duolingo courses because not all Duolingo courses are at the same level of content development and CEFR alignment.

We acknowledge possible concerns regarding sample size when conducting factor analysis. The literature reports wide variation in minimum sample size recommendations which vary from two to 20 participants per variable (Pituch & Stevens, 2015). In any case, there is a paucity of empirical support for these guidelines, especially in terms of subjects-to-variables and absolute ranges (Mundfrom et al., 2005). In fact, some researchers argue that factors such as the robustness of factor loadings (factor saturation) and the number of variables demonstrating high loadings are more crucial than sample size when assessing the adequacy of sample sizes in factor analysis (Guadagnoli & Velicer, 1988). While sample size concerns are valid, our exploratory data shows compelling factor loadings across many variables, lending confidence to the identification of patterns within the confines of our dataset. Thus, we feel that the use of factor analysis in Research Question 2 was appropriate as a preliminary step to inform future, more extensive studies.

Also, while learner proficiency gains were assessed by a widely-used standardized test (STAMP 4S), the instruments used to measure pre- and post-test gains of the other linguistic measures were adapted from previously used researcher-developed tests. This is common practice in SLA studies. Indeed, the EIT and vocabulary tests upon which we based our own measures were themselves iterations of previous versions of tests. For the present study, revalidation of each of these measures was deemed impractical given the study's limited scale and scope. While we consider the adaptations made to the source materials to be reasonable, appropriate, and pragmatic, we make no claims regarding the suitability of these instruments for use in other learning contexts.

Building upon this study, future studies may wish to explore several additional areas. First, this study focused on examining the learning outcomes of participants with low Spanish proficiency during the pre-test phase. As Duolingo continues to develop more advanced course content, future research should also include learners with higher proficiency levels in order to assess its effectiveness. Second, this study followed a within-subject quasi-experimental design. Future studies may consider a between-subject design, which would allow for comparisons between Duolingo and other instructional contexts, including learner use of other language learning tools. Finally, though some learners reported that this learning experience made them more competent and confident in real world interactions in the target language, this study didn't assess the practical application of Spanish in real communication scenarios. Future research should examine how learners apply their language skills in authentic real-world tasks.

## Pedagogical Implications

The study reveals that as little as 2.25 hours a week of self-directed learning with Duolingo over 12 weeks can lead to substantial gains in overall language proficiency and specific skill area linguistic knowledge, and these gains are reflected in both receptive and productive skills. While Duolingo courses primarily focus on sentence-level vocabulary and grammar, incorporating some longer-form content, the current study suggests that this seemingly discrete vocabulary and grammar knowledge can be applied to integrative tasks targeting both receptive and productive skills, as indicated by Loewen et al. (2020). A key factor in this success may be Duolingo's gamified approach to language learning, which seems to make the experience more engaging and motivating.

Duolingo's CEFR-aligned courses aim to teach vocabulary and grammar through theme-based communication functions (Freeman et al., 2023). The significant correlation between the number of sessions completed and language knowledge gains underscores the effectiveness of the course content in facilitating mastery of the material. Additionally, the session accuracy rate strongly predicts proficiency improvements, suggesting that learners making fewer mistakes experience greater enhancements in target language proficiency.

## Conclusion

This study offers methodologically sound and compelling evidence that learning Spanish with Duolingo for a short time each day over three months significantly improves both overall proficiency and specific skill area ability for both receptive and productive skills for learners at an intermediate-low level or lower. By accounting for learners' prior proficiency with a pre-test and using a third-party standardized test along with a set of researcher-created language assessments, we can confidently assess the efficacy of Duolingo for overall proficiency and skill area development. The combined and robust evidence of efficacy demonstrates that Duolingo's Spanish course for English speakers not only assists learners in mastering course content but also fosters overall Spanish proficiency.

## Acknowledgements

This study was supported financially by Duolingo. The authors gratefully acknowledge the assistance of Mariana Lara Albert, Chukwudi Obite, Dr. Bozena Pajak, David Parks, Dr. Lucy Skidmore, Dr. Olena Tanchyk, and Giannina van Straaten in this study.

## Notes

[1] For a description of the psychometric properties of the STAMP 4S Spanish Test, please refer to Avant (2023, December). *Research*. https://avantassessment.com/data-analysis/research

[2] In order to avoid a test/retest effect, we followed ACTFL’s policy that does not generally permit test takers to retake an assessment within 90 days of a test administration (ACTFL, 2024).

[3] Possible reasons for receiving a NR (Not Ratable) grade, as opposed to a score of zero, include technical issues as well as other issues related to background noises, which prevent the rater from adequately hearing the response. The Avant description of “Possible Reasons for NR (Not Ratable) Status” can be found here https://avantassessment.com/guides/coordinator/reporting/stamp-ws#PossibleNR

[4] Session accuracy refers to the average accuracy rate of all the sessions learners engaged in on Duolingo. Session accuracy rate was found to be significantly correlated with learners’ language

proficiency in Jiang, Hopman, et al. (2024).

5. Item 1 “How much Spanish do you know now?” was reported on a 10-point scale and later converted.

6. We initially used the Factor Analysis class from scikit-learn with 3 factors but questions loaded onto a single factor, possibly misrepresenting underlying constructs. Switching to Factor Analyzer improved control (we employed varimax rotation), resulting in distinct, interpretable factors with questions loading appropriately. The Principal Axis extraction method was employed.

## References

ACTFL. (2024). *Quality control of ACTFL assessments*. https://www.actfl.org/assessments/quality-control-of-actfl-assessments

Avant. (2023, December). *Research*. https://avantassessment.com/data-analysis/research

Brown, K. (2023, June 7). *CALL as a driver of equity, access, and outcomes in an increasingly connected world* [Opening plenary]. CALICO Annual Conference, University of Minnesota, Minneapolis, MN, United States.

Burston, J., & Athanasiou, A. (2020). Twenty-five years of MALL experimental implementation studies: What do we really know about it? In A. Andujar (Ed.), *Recent tools for computer- and mobile-assisted foreign language learning* (pp. 35-59). IGI Global. https://doi.org/10.4018/978-1-7998-1097-1.ch002

Burston, J., & Giannakou, K. (2022). MALL language learning outcomes: A comprehensive meta-analysis 1994–2019. *ReCALL, 34*(2), 147–168. https://doi.org/10.1017/S0958344021000240

Byrne, J., Ito, T., & Furuyabu, M. (2022). Digitally nudged learning: A nudged gamification study intervention. *International Journal of Emerging Technologies in Learning, 17*(12), 42–60. https://doi.org/10.3991/ijet.v17i12.30567

Chen, Z., Chen, W., Jia, J., & An, H. (2020). The effects of using mobile devices on language learning: A meta-analysis. *Educational Technology Research and Development, 68*(4), 1769–1789. https://doi.org/10.1007/s11423-020-09801-5

Clément, R., Dörnyei, Z., & Noels, K. A. (1994). Motivation, self-confidence, and group cohesion in the foreign language classroom. *Language Learning, 44*(3), 417–448. https://doi.org/10.1111/j.1467-1770.1994.tb01113.x

Collentine, J., & Freed, B. F. (2004). Learning context and its effects on second language acquisition: Introduction. *Studies in Second Language Acquisition, 26*(2), 153–171. https://doi.org/10.1017/S0272263104262015

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

Curry, D. (2023, May 15). Language learning app revenue and usage statistics. *Business of Apps*. https://www.businessofapps.com/data/language-learning-app-market

East, M., & King, C. (2012). L2 learners’ engagement with high stakes listening tests: Does technology have a beneficial role to play? *CALICO Journal, 29*(2), 208–223. https://doi.org/10.11139/cj.29.2.208-223

Freeman, C., Kittredge, A., Wilson, H., & Pajak, B. (2023). *The Duolingo method for app-based teaching and learning* [Duolingo Research Report]. Retrieved on September 8, 2023 from https://duolingo-papers.s3.amazonaws.com/reports/duolingo-method-whitepaper.pdf

Gardener, R. C. (1985). *Social psychology and second language learning: The role of attitudes and motivation*. Edward Arnold.

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin, 103*(2), 265–275. https://doi.org/10.1037/0033-2909.103.2.265

He, X., & Loewen, S. (2022). Stimulating learner engagement in app-based L2 vocabulary self-study: Goals and feedback for effective L2 pedagogy. *System, 105*, 1–13. https://doi.org/10.1016/j.system.2021.102719

Izura, C., Cuetos, F., & Brysbaert, M. (2014). Lextale-Esp: A test to rapidly and efficiently assess the Spanish vocabulary size. *Psicológica, 35*(1), 49–66.

Jiang, X., Hopman, E., Revuni, B., & Kittredge, A. (2024). Duolingo path meets expectations for proficiency outcomes [Duolingo Research Report]. Retrieved from https://duolingo-papers.s3.amazonaws.com/reports/Duolingo_whitepaper_language_read_listen_write_speak_2024.pdf

Jiang, X., Peters, R., Plonsky, L., & Pajak, B. (2024). The effectiveness of Duolingo English courses in developing reading and listening proficiency. *CALICO Journal, 41*(3). https://doi.org/10.1558/cj.26704

Jiang, X., Rollinson, J., Plonsky, L., Gustafson, E., & Pajak, B. (2021). Evaluating the reading and listening outcomes of beginning-level Duolingo courses. *Foreign Language Annals, 54*(4), 974–1002. https://doi.org/10.1111/flan.12600

Kessler, M. (2021). Supplementing mobile-assisted language learning with reflective journal writing: A case study of Duolingo users' metacognitive awareness. *Computer Assisted Language Learning, 36*(5-6), 1040–1063. https://doi.org/10.1080/09588221.2021.1968914

Kessler, M., Loewen, S., & Gönülal, T. (2023). Mobile-assisted language learning with Babbel and Duolingo: Comparing L2 learning gains and user experience. *Computer Assisted Language Learning*, 1–25. https://doi.org/10.1080/09588221.2023.2215294

Kohnke, L. (2020). Exploring learner perception, experience and motivation of using a mobile app in L2 vocabulary acquisition. *International Journal of Computer-Assisted Language Learning, 10*(1), 15–26. https://doi.org/10.4018/IJCALLT.2020010102

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

Li, J. (2018). Digital affordances on WeChat: Learning Chinese as a second language. *Computer Assisted Language Learning, 31*(1–2), 27–52. https://doi.org/10.1080/09588221.2017.1376687

Lin, C.-H., & Warschauer, M. (2015). Online foreign language education: What are the proficiency outcomes? *The Modern Language Journal, 99*(2), 394–397. https://doi.org/10.1111/modl.12234_1

Loderer, K., Pekrun, R., & Lester, J. C. (2020). Beyond cold technology: A systematic review and meta-analysis on emotions in technology-based learning environments. *Learning and Instruction, 70.* https://doi.org/10.1016/j.learninstruc.2018.08.002

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. 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*(2), 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. https://doi.org/10.1111/modl.12234_3

Lord, G. (2016). Rosetta Stone for language learning: An exploratory study. *IALLT Journal of Language Learning Technologies*, *46*(1), 1–35. https://doi.org/10.17161/iallt.v46i1.8552

M Science. (2023). *DUOL Q2'23 first look*. M Science. Retrieved on May 12, 2023 from https://mscience.com/

Masgoret, A.-M., & Gardner, R. C. (2003). Attitudes, motivation, and second language learning: A meta-analysis of studies conducted by Gardner and Associates. *Language Learning, 53*(1), 123–163. https://doi.org/10.1111/1467-9922.00212

Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing, 5*(2), 159–168. https://doi.org/10.1207/s15327574ijt0502_4

OpenAI. (2023). *ChatGPT* (Mar 14 version). [Large language model]. https://www.openai.com

Park, H. I., Solon, M., Dehghan-Chaleshtori, M., & Ghanbar, H. (2022). Proficiency reporting practices on second language acquisition research: Have we made any progress? *Language Learning, 72*(1), 198–236. https://doi.org/10.1111/lang.12475

Peng, H., Jager, S., & Lowie, W. (2021). Narrative review and meta-analysis of MALL research on L2 skills. *ReCALL, 33*(3), 278–295. https://doi.org/10.1017/S0958344020000221

Pituch, K. A., & Stevens, J. P. (2015). *Applied multivariate statistics for the social sciences: Analyses with SAS and IBM's SPSS* (6th ed.). Routledge.

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*(4), 878–912. https://doi.org/10.1111/lang.12079

Rachels, J. R., & Rockinson-Szapkiw, A. J. (2017). The effects of a mobile gamification app on elementary students' Spanish achievement and self-efficacy. *Computer Assisted Language Learning, 31*(1-2), 72–89. https://doi.org/10.1080/09588221.2017.1382536

Rosell-Aguilar, F. (2018). Autonomous language learning through a mobile application: A user evaluation of the *busuu* app. *Computer Assisted Language Learning, 31*(8), 854–881. https://doi.org/10.1080/09588221.2018.1456465

Santos, V. (2022, June). *The STAMP 4S English test reading and listening sections* [Technical report]. Avant Assessment.

Shortt, M., Tilak, S., Kuznetcova, I., Martens, B., & Akinkuolie, B. (2023). Gamification in mobile-assisted language learning: A systematic review of Duolingo literature from public release of 2012 to early 2020. *Computer Assisted Language Learning, 36*(3), 517–554. https://doi.org/10.1080/09588221.2021.1933540

Solon, M., Park, H. I., Henderson, C., & Dehghan-Chaleshtori, M. (2019). Revisiting the Spanish elicited imitation task: A tool for assessing advanced language learners? *Studies in Second Language Acquisition, 41*(5), 1027–1053. https://doi.org/10.1017/S0272263119000342

Sporn, Z., Chanter, J., & Meehan, D. (2020). Babbel language learning podcasts. *International Journal of Advanced Corporate Learning, 13*(3), 43–49. https://doi.org/10.3991/ijac.v13i3.17025

Stockwell, G. (2022). *Mobile assisted language learning: Concepts, contexts and challenges*. Cambridge University Press. https://doi.org/10.1017/9781108652087.001

Sudina, E., & Plonsky, L. (2024). The effects of frequency, duration, and intensity on L2 learning through Duolingo: A natural experiment. *Journal of Second Language Studies, 7*(1), 1–43. https://doi.org/10.1075/jsls.00021.plo

Tarone, E. (2015). Online foreign language education: What are the proficiency outcomes? *The Modern Language Journal*, *99*(2), 392–393. https://doi.org/10.1111/modl.12220

*The Common European Framework of Reference for Languages (CEFR)*. (n.d.). Cambridge International Education. Retrieved from https://help.cambridgeinternational.org/hc/en-gb/articles/115004446685-The-Common-European-Framework-of-Reference-for-Languages-CEFR/

Thomas, M. (2006). Research synthesis and historiography: The case of assessment of second language proficiency. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 279–298). John Benjamins.

Vesselinov, R. (2009). *Measuring the effectiveness of Rosetta Stone* [White paper]. Retrieved on February 4, 2024, from https://resources.rosettastone.com/CDN/us/pdfs/Measuring_the_Effectiveness_RS-5.pdf

Vesselinov, R., & Grego, J. (2012). *Duolingo effectiveness study* [White paper]. Retrieved on February 4, 2024, from https://www.languagezen.com/pt/about/english/Duolingo_Efficacy_Study.pdf

Vesselinov, R., & Grego, J. (2016a). *The Babbel efficacy study* [White paper]. Retrieved on February 4, 2024, from http://comparelanguageapps.com/documentation/Babbel2016study.pdf

Vesselinov, R., & Grego, J. (2016b). *The busuu efficacy study* [White paper]. Retrieved on February 4, 2024, from http://comparelanguageapps.com/documentation/The_busuu_Study2016.pdf

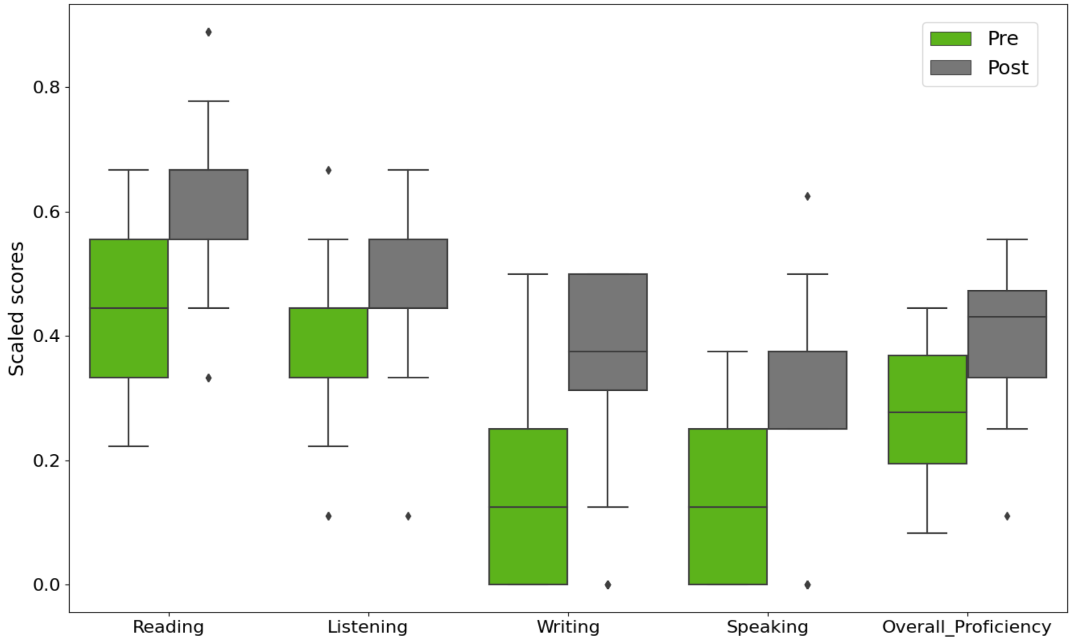
Vesselinov, R., & Grego, J. (2018). *The italki efficacy study* [White paper]. Retrieved on February 4, 2024, from https://www.gettingsmart.com/wp-content/uploads/2018/01/italki2018FinalReport.pdf

Winke, P., Gass, S. M., Soneson, D., Rubio, F., & Hacking, J. F. (2020). Foreign language proficiency test data from three American universities, [United States], 2014-2017. *Inter-university Consortium for Political and Social Research*. https://doi.org/10.3886/ICPSR37499.V1

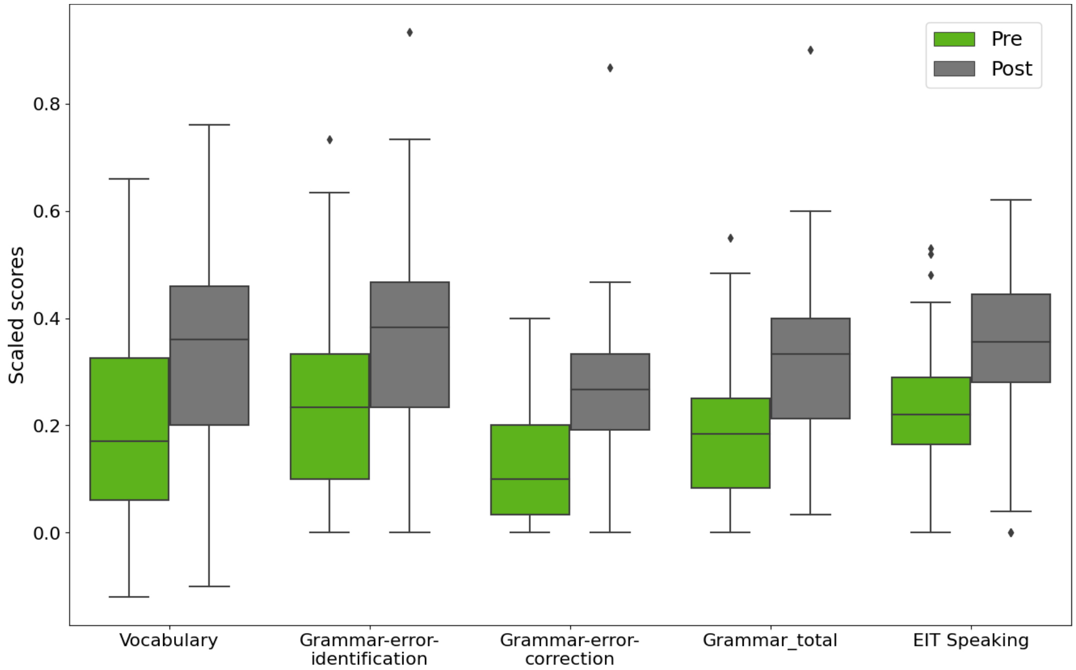
Yu, Z., Xu, W., & Sukjairungwattana, P. (2023). Motivation, learning strategies, and outcomes in mobile English language learning. *The Asia-Pacific Education Researcher, 32*, 545–560. https://doi.org/10.1007/s40299-022-00675-0

## Appendix A. Comparison of Pre- and Post-test Scores across Proficiency and Linguistics Measures le of Appendix

### *Proficiency*



### *Linguistic*



*Note.* These scaled scores reflect the normalized scores of each test based on the predefined min and max values, which varied. This scaling process transforms the scores so that they fall within a 0 to 1 range, where 0 corresponds to the minimum test score and 1 corresponds to the maximum test score for each test type. For example, the Reading and Listening scores on the STAMP 4S Spanish test are on a 1-9 scale, whereas Writing and Speaking are scored on a 1-8 scale.

## Appendix B. User Experience Factors and Item Loadings

| Question # | | h² | **Factor 1** | **Factor 2** | **Factor 3** |
|---|---|---|---|---|---|
| | | | **Perceived Efficacy** | **Continued Engagement** | **Positive User Experience** |
| 7.9 | I feel more confident using Spanish after 12 weeks learning on Duolingo. | 0.85 | **0.85** | 0.27 | 0.24 |
| 7.11 | I feel more comfortable speaking Spanish after 12 weeks learning on Duolingo. | 0.78 | **0.79** | 0.18 | 0.35 |
| 7.15 | I can use the language I learned on Duolingo in practical situations. | 0.72 | **0.75** | 0.29 | 0.28 |
| 7.10 | I am surprised how much I learned after 12 weeks with Duolingo. | 0.83 | **0.73** | 0.36 | 0.40 |
| 7.1 | From my experience, Duolingo is an effective way to learn a language. | 0.78 | **0.72** | **0.51** | 0.02 |
| 7.2 | From my experience, Duolingo is an efficient/fast way to learn a language. | 0.73 | **0.70** | 0.48 | -0.02 |
| 7.16 | Using Duolingo makes me confident in my ability to learn new languages. | 0.75 | **0.67** | 0.13 | **0.53** |
| 7.3 | From my experience, Duolingo provides lasting language skills. | 0.69 | **0.65** | 0.46 | 0.21 |
| 7.12 | I feel more comfortable reading Spanish after 12 weeks learning on Duolingo. | 0.61 | **0.61** | 0.09 | 0.48 |
| 7.13 | I feel more comfortable listening to Spanish after 12 weeks learning on Duolingo. | 0.61 | **0.59** | 0.34 | **0.51** |
| 6 | To what extent do you believe learning apps can teach languages? | 0.67 | **0.59** | **0.54** | 0.19 |
| 3 | How likely are you to continue studying Spanish? | 0.75 | 0.24 | **0.81** | 0.19 |

| | | | | | |
|---|---|---|---|---|---|
| 4 | How likely are you to continue learning Spanish on Duolingo? | 0.74 | 0.32 | **0.76** | 0.23 |
| 2 | How interested are you in learning Spanish? | 0.60 | 0.08 | **0.76** | 0.13 |
| 7.5 | From my experience, learning on Duolingo is a good use of my time. | 0.66 | 0.29 | **0.65** | 0.40 |
| 7.4 | From my experience, Duolingo is a fun way to learn a language. | 0.77 | 0.24 | 0.45 | **0.71** |
| 7.8 | I enjoy learning Spanish on Duolingo. | 0.87 | 0.31 | **0.60** | **0.65** |
| 7.14 | I am more motivated in learning Spanish after 12 weeks on Duolingo. | 0.76 | 0.32 | **0.55** | **0.60** |
| 7.6 | I would be excited to tell my friends about Duolingo. | 0.69 | 0.43 | 0.44 | **0.56** |
| 5 | How likely are you to learn (an)other language(s) on Duolingo? | 0.32 | 0.11 | 0.06 | **0.55** |
| 7.7 | I’m proud to be seen using Duolingo. | 0.41 | 0.29 | 0.32 | 0.47 |
| 1 | How much Spanish do you know now? | 0.08 | 0.24 | 0.05 | 0.13 |

*Note.* Loadings of 0.5 mean that 25% (0.5^2) of the variance in the survey question is explained by the factor. The varimax rotation method employed here does not change the interpretation of factor loadings in terms of the proportion of variance explained. Varimax rotation is an orthogonal rotation method that aims to simplify the factor structure by maximizing the variance of the squared loadings within factors. It makes the factor loadings more distinct and easier to interpret, but it does not change the overall proportion of variance explained by the factors.

## Appendix C. Average Ratings and SDs of Survey Items by Factor

| Question # | | M | SD |
|---|---|---|---|
| | Factor 1: Perceive learning efficacy | | |
| 7.9 | I feel more confident using Spanish after 12 weeks learning on Duolingo. | 3.60 | 1.30 |
| 7.11 | I feel more comfortable speaking Spanish after 12 weeks learning on Duolingo. | 3.19 | 1.36 |
| 7.15 | I can use the language I learned on Duolingo in practical situations. | 3.29 | 1.18 |
| 7.10 | I am surprised how much I learned after 12 weeks with Duolingo. | 3.63 | 1.27 |
| 7.1 | From my experience, Duolingo is an effective way to learn a language. | 3.73 | 1.03 |
| 7.2 | From my experience, Duolingo is an efficient/fast way to learn a language. | 3.44 | 1.20 |
| 7.16 | Using Duolingo makes me confident in my ability to learn new languages. | 3.48 | 1.34 |
| 7.3 | From my experience, Duolingo provides lasting language skills. | 3.46 | 1.18 |
| 7.12 | I feel more comfortable reading Spanish after 12 weeks learning on Duolingo. | 3.60 | 1.28 |
| 7.13 | I feel more comfortable listening to Spanish after 12 weeks learning on Duolingo. | 3.21 | 1.32 |
| 6 | To what extent do you believe learning apps can teach languages? * | 3.60 | 0.94 |
| | Factor 2: Continued engagement | | |
| 7.5 | From my experience, learning on Duolingo is a good use of my time. | 4.08 | 1.03 |
| 3 | How likely are you to continue studying Spanish? * | 3.58 | 1.21 |
| 4 | How likely are you to continue learning Spanish on Duolingo? * | 3.73 | 1.31 |
| 2 | How interested are you in learning Spanish? * | 3.92 | 1.02 |
| 7.1 | From my experience, Duolingo is an effective way to learn a language. | 3.73 | 1.03 |
| 6 | To what extent do you believe learning apps can teach languages? * | 3.25 | 1.18 |
| 7.8 | I enjoy learning Spanish on Duolingo. | 3.90 | 1.21 |
| 7.14 | I am more motivated in learning Spanish after 12 weeks on Duolingo. | 3.71 | 1.29 |
| | Factor 3: Positive user experience | | |

| | | | |
|---|---|---|---|
| 7.4 | From my experience, Duolingo is a fun way to learn a language. | 4.10 | 1.15 |
| 7.8 | I enjoy learning Spanish on Duolingo. | 3.90 | 1.21 |
| 7.14 | I am more motivated in learning Spanish after 12 weeks on Duolingo. | 3.71 | 1.29 |
| 7.6 | I would be excited to tell my friends about Duolingo. | 3.65 | 1.28 |
| 7.7 | I'm proud to be seen using Duolingo. | 3.52 | 1.18 |
| 5 | How likely are you to learn (an)other language(s) on Duolingo? * | 3.08 | 1.46 |
| 7.16 | Using Duolingo makes me confident in my ability to learn new languages. | 3.48 | 1.34 |
| 7.13 | I feel more comfortable listening to Spanish after 12 weeks learning on Duolingo. | 3.21 | 1.32 |

*Note.* * designates items that originally used a 0-5 scale. To make these scores comparable to the items using a 1-5 scale, a linear scaling transformation was applied (see Han, 2023).

## Appendix D. Main Themes and Mention Frequency in Open-Ended Responses

**Question 9:**
When learning a language on an app, how do you know if the app you're using is effective?

| | Theme | Mentions |
|---|---|---|
| 1. | Language Proficiency and Progress | 32 |
| 2. | Application and Real-world Interaction | 22 |
| 3. | Engagement and Challenge | 5 |
| 4. | Feedback and Performance Monitoring | 4 |

**Question 10:**
What made you feel that Duolingo is helpful or not helpful for improving your language skills?

| | Theme | Mentions |
|---|---|---|
| 1. | Language Skills Development | 19 |
| 2. | Learning Experience | 16 |
| 3. | Practical Language Use | 12 |
| 4. | Content and Curriculum | 4 |

**Question 11:**
How would you describe your experience learning Spanish on Duolingo?

| | Theme | Mentions |
|---|---|---|
| 1. | User Experience & Accessibility | 42 |
| 2. | Content Interaction & Learning Process | 10 |
| 3. | Learning Outcomes & Application | 10 |
| 4. | Community & Social Learning | 4 |

**Question 12:**
How did it feel to use Duolingo the way you did?

| | Theme | Mentions |
|---|---|---|
| 1. | Engagement & Motivation | 43 |
| 2. | Practicality & Learning Outcomes | 13 |
| 3. | Habit Formation & Routine | 6 |
| 4. | Learning Experience | 1 |

**Question 13:**
What do you like most about Duolingo?

| | Theme | Mentions |
|---|---|---|
| 1. | Learning Efficacy and Tools | 21 |
| 2. | User Experience and Design | 14 |
| 3. | Engagement and Gamification | 11 |
| 4. | Social Connectivity and Competition | 4 |

**Question 14:**
What would you like Duolingo to change?

| | Theme | Mentions |
|---|---|---|
| 1. | Content Enrichment and Diversification | 14 |
| 2. | Interface and Interaction Enhancements | 13 |
| 3. | Grammar and Communication Focus | 10 |
| 4. | Customization and User Agency | 7 |

**Question 15:**
Is there anything else you would like to say about your experience learning on Duolingo?

| | Theme | Mentions |
|---|---|---|
| 1. | Positive Learning Experience | 11 |
| 2. | Mixed Feedback on Content | 4 |
| 3. | Desire for More Interactive Features | 4 |
| 4. | Suggestions for Improvement | 3 |

## About the Authors

Bryan Smith is Professor of Applied Linguistics and CALL at Arizona State University. His research, which focuses on the intersection of CALL and SLA, has appeared in various applied linguistics and CALL journals. He is the co-editor of the *CALICO Journal*. Bryan Smith is the corresponding author.
**E-mail:** bryansmith@asu.edu
**ORCiD:** https://orcid.org/0000-0002-9813-6365

Xiangying Jiang is a lead learning scientist at Duolingo where she works on learning assessment and efficacy research. She holds a PhD in Applied Linguistics (Northern Arizona University). Her work has appeared in journals such as CALICO Journal, The Modern Language Journal, Reading Psychology, and Foreign Language Annals.
**E-mail:** xiangying@duolingo.com

Ryan Peters is a Senior Learning Scientist at Duolingo with a PhD in Speech, Language, and Hearing Sciences from Purdue University. His research has largely focused on attention and learning in first and second language acquisition.
**E-mail:** ryanpeters@duolingo.com