

ARTICLE



## The effectiveness of app-based and classroom-based instruction on L2 learning and motivation

Beatriz González-Fernández\*, University of Sheffield

Inés de la Viña, University of Nottingham

### Abstract

*Language learners around the world are increasingly using applications (apps) to learn second/foreign languages (L2). However, research on the effectiveness of these apps for developing general language proficiency, particularly compared to classroom-based instruction, is still limited. This study examined the L2-English proficiency and lexical development of 337 L1-Spanish learners enrolled in either app-based (Duolingo) or classroom-based instruction over a 16-week period. Using a pretest-posttest design, participants completed a background and motivation questionnaire, a general L2-English proficiency test, and two vocabulary tests tapping into receptive and productive knowledge. Results showed that both modes of instruction led to significant language gains. Duolingo learners outperformed classroom learners on measures of general L2 proficiency and receptive vocabulary, while classroom learners showed significantly greater improvement in listening skills. Gains in productive vocabulary knowledge were comparable across both groups. Participants in both groups also reported generally high levels of L2 motivation throughout the study, with Duolingo learners indicating slightly higher levels of interest in the course. Overall, these findings suggest that app-based learning can support certain aspects of L2 development, particularly receptive grammar and vocabulary knowledge, while classroom-based instruction is more beneficial for developing listening skills.*

**Keywords:** app-based learning; second language proficiency; vocabulary knowledge; motivation

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

**APA Citation:** González-Fernández, B. & de la Viña, I. (2025). The effectiveness of app-based and classroom-based instruction on L2 learning and motivation. *Language Learning & Technology*, 29(1), 1–18. <https://doi.org/10.64152/10125/73656>

### Introduction

Computer and mobile-assisted language learning (CALL and MALL respectively) tools, particularly language learning applications (apps), have become increasingly popular among learners worldwide (Burston & Giannakou, 2022; Loewen et al., 2019). This surge in instructional technologies has transformed the way in which L2s are learnt, significantly impacting the field of (instructed) second language acquisition (SLA) in and outside the classroom. Given their many benefits for L2 learning and instruction, such as *autonomy* of study, *flexibility* of use and *individualization* of learning, the use of MALL technologies, especially language learning apps, is predicted to continue thriving (Loewen, 2020).

Yet, despite the popularity and rapid expansion of language learning apps, research investigating their effectiveness in promoting L2 proficiency development is lagging behind adoption (Rachels & Rockinson-Szapkiw, 2018). Consequently, SLA researchers and teachers continue to debate the role of apps in language learning and how they can complement and assist classroom-based instruction (Loewen et al., 2020; Shortt et al., 2023). The present study aims to shed light on this issue by exploring the effectiveness of app-based learning in facilitating the L2-English proficiency and lexical development of L1-Spanish learners and comparing it to the linguistic development of similar learners enrolled in face-to-face classroom-based instruction<sup>1</sup>.

\* **Corresponding Author:** Beatriz González-Fernández, [b.gonzalez-fernandez@sheffield.ac.uk](mailto:b.gonzalez-fernandez@sheffield.ac.uk)

## Literature Review

### Mobile-Assisted Language Learning and Instructed SLA

Alongside classroom-based instruction, the use of CALL and MALL technologies is now considered one of the core contexts of instructed SLA (Loewen, 2020). These technologies have been effectively employed as complements to classroom-based instruction both during class time and outside the class. Guaqueta and Castro-Garces (2018) integrated the use of Duolingo into English lessons delivered in a conventional classroom setting over 6 months and observed enhanced vocabulary knowledge and improved attitudes toward language learning. Wu (2015) found that combining in-person English lessons with the autonomous use of a vocabulary-building app used outside the classroom capitalized on “dead time”, such as commuting, resulting in increased study time.

Apps’ growing popularity lies in offering a convenient and affordable path to L2 learning (Rachels & Rockinson-Szapkiw, 2018). App-based learning is not intended to replace teacher-led instruction and instead offers alternative and complementary means to L2 learning, particularly when access to formal educational environments is limited (Loewen, 2020). However, for many learners, apps now function as the primary or only method of autonomous language study (Loewen et al., 2020). Since the type of instruction learners receive influences their L2 development greatly (Norris & Ortega, 2000), it is crucial to examine the effectiveness of independent app-based instruction relative to classroom-based instruction to advance our knowledge of MALL apps’ role in instructed SLA.

### Effectiveness of MALL Apps

As MALL apps gain popularity, understanding their efficacy in promoting L2 learning becomes essential for both researchers and practitioners (Burston & Giannakou, 2022). While most research on the use of MALL apps has focused on describing their design instead of investigating their influence in L2 development (Shortt et al., 2023), there has been in recent years a rise in studies examining the acquisition of L2s through language learning apps (e.g., Jiang et al., 2024; Loewen et al., 2019, 2020; Sudina & Plonsky, 2024).

The majority of MALL studies conclude that apps are effective learning tools for developing L2 competence (Burston & Giannakou, 2022). Significant learning gains have been reported particularly regarding reading ability (e.g., Jiang et al., 2024) and knowledge of vocabulary and grammar at the receptive level (e.g., Loewen et al., 2020; Rachels & Rockinson-Szapkiw, 2018). However, apps’ effectiveness in developing oral skills and productive knowledge of lexis and grammar is far less evident. Most studies do not focus on these language aspects (Shortt et al., 2023), and when they do, results show small (or even a lack of) significant gains in listening skills (e.g., Jiang, Rollinson, Plonsky, et al., 2021), speaking skills (e.g., Loewen et al., 2019, 2020; Lord, 2015), or productive vocabulary knowledge. In a meta-analysis on app efficacy for vocabulary development, Lin and Lin (2019) found that only 7 of the 29 studies examined focused on productive vocabulary, which showed smaller effect sizes than receptive vocabulary. Similarly, Jiang, Rollinson, Chen, et al., (2021) found that productive vocabulary knowledge scored lowest among their L2 measures. More research on app-based learning is thus needed, particularly in underexplored areas like productive vocabulary and oral skills.

In addition, few studies have directly compared the effectiveness of app-based and classroom-based L2 instruction (Jiang, Rollinson, Plonsky, et al., 2021). One early example is Lord (2015), who compared beginning-level Spanish learners in classroom-based instruction with those using the app Rosetta Stone to learn English over 16 weeks. She found no differences between the two groups’ performance on standardized L2-English tests, despite the app-based group dedicating substantially less time in the program. Yet, she noted that students learning through the app faced more challenges in conversation compared to their classroom-based counterparts. Rachels and Rockinson-Szapkiw (2018) showed that L2-Spanish primary-school students using Duolingo achieved similar grammar and vocabulary gains to classroom learners when study time was equal (40 minutes a week for 12 weeks). More recently, Jiang,

Rollinson, Plonsky, et al. (2021) showed that Duolingo learners of two different L2s (Spanish and Portuguese) attained comparable reading and listening skills to university students studying the L2 for four semesters, while spending only half the time on program.

Prior research offers mixed findings on the effectiveness of app-based versus classroom instruction. Some studies report comparable outcomes in reading and listening with fewer study hours for app users (Jiang et al., 2021b), while others highlight stronger oral skills among classroom learners despite similar test performance (Lord, 2015). Other studies find both methods equally effective, particularly in enhancing lexical and grammatical knowledge (Rachels & Rockinson-Szapkiw, 2018).

The seemingly inconsistent findings across studies might be explained by their limitations, which include: a) lack of pre-test data and control/comparison group (e.g., Jiang, Rollinson, Plonsky, et al., 2021; Loewen et al., 2019, 2020); b) lack of or minimal control of study time by participants (e.g., Lord, 2015); c) small variability in learning environment and samples (i.e., mainly US university learners); and d) small sample sizes (e.g., Lord, 2015,  $n=12$ ). These limitations highlight the need for further research to compare app-based vs. classroom-based instruction for L2 development (Sudina & Plonsky, 2024), and the present study aims to address this issue and the previous limitations.

### **Motivation and Language Learning**

While the effect of individual differences is well-established in SLA, their impact on MALL remains unclear. Learner factors such as motivation or engagement with the course can predict language development in both app-based (He & Loewen, 2022; Loewen et al., 2020) and classroom instruction (Saito et al., 2018). Learning a language can be challenging and stressful, is time-consuming and requires perseverance to keep studying and practicing (Shortt et al., 2023). Without adequate motivation and engagement with the language, students are less likely to succeed. These two concepts are interlinked: L2 motivation drives active learning, whereas engagement is the behavioral manifestation of motivation (Dörnyei, 2019).

App-based learners have been found to struggle with motivation, engagement, and persistence using apps over the long term, affecting their learning outcomes. García Botero et al. (2019) found that university students using Duolingo to supplement their instruction logged in fewer than 10 times in a year. Similarly, Loewen et al. (2019) reported only 22% of learners met a 34-hour semester goal, affecting their motivation and persistence.

Conversely, some scholars argue that apps' engaging designs, attractive interface and gamified features (i.e., leaderboards, streaks) may enhance motivation and engagement more than classroom-based instruction (Dehganzadeh & Dehganzadeh, 2020; He & Loewen, 2022; James & Mayer, 2019). James and Mayer (2019) examined college students learning L2-Italian at home using Duolingo versus learning it using an online slideshow during 7 sessions. They found that, while the groups did not differ on linguistic achievement, Duolingo learners reported the experience to be more enjoyable, appealing and less difficult, as well as more willingness to continue studying the language. Dehganzadeh and Dehganzadeh (2020) found that gamified language learning boosts motivation, engagement, and persistence, highlighting the motivational value of MALL apps. The unique features of app-based self-study and classroom-based settings may influence learners' motivation and engagement differently, but it remains unclear how and why, highlighting the need for research to examine this issue (He & Loewen, 2022).

### **Duolingo Course**

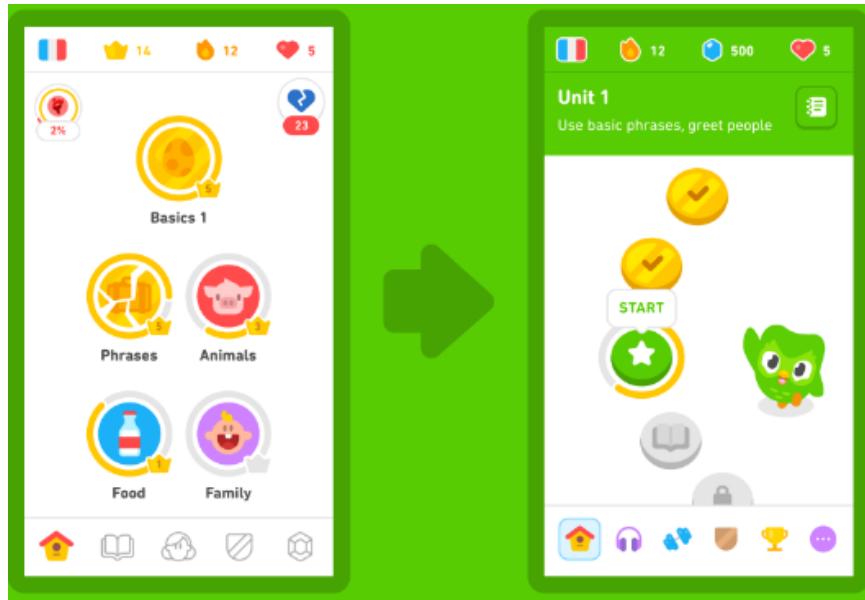
Duolingo is a leading MALL app and the most investigated platform by SLA researchers (Dehganzadeh & Dehganzadeh, 2020). This study examines the effectiveness of Duolingo's most popular course – English for Spanish speakers, with over 44 million learners – and compares it to a classroom-based course. Both courses align with the CEFR (Council of Europe, 2017). Our participants were finishing the A1-level content (beginner, section A1.2) and starting A2 (elementary) at the pretest time. Based on Duolingo data (personal communication), learners were expected to progress through A2 during the 16-week

period. Lessons cover communicative topics (e.g., family, food, travel) and teach vocabulary and grammar via translation, multiple-choice, and gap-filling exercises. Duolingo also includes gamified features such as rewards and user rankings (Shortt et al., 2023).

This study used Duolingo Version 2 (2022), which requires all users to follow a linear path through different difficulty levels, similar to a classroom syllabus (Figure 1), unlike earlier versions where learners navigated the platform autonomously and could skip content (Jiang, Rollinson, Plonsky, et al., 2021). This updated design ensures more consistent progress and better comparability with classroom instruction.

**Figure 1**

*Duolingo Version 2 Learning Path*



The classroom course in this study also aligns with the CEFR. Classroom learners were enrolled in a face-to-face A2-level English course, having completed the A1 level in the same mode. Unlike Duolingo, the classroom syllabus used English as the main language of instruction, emphasized L2 use and social interaction, and discouraged L1 use. Lessons focused on spoken interaction, listening, and everyday conversations. Duolingo's gamified learning approach further distinguishes it from classroom-based instruction.

### **The Present Study**

Expanding on previous MALL research, this pretest-posttest study compares Spanish speakers' L2-English proficiency and lexical development after learning through Duolingo or classroom-based instruction. It also evaluates learners' sustained motivation and engagement, while controlling for instruction length (16 weeks) and weekly study time (3-4 hours per week) for both groups. The following research questions (RQ) are addressed:

RQ1. How effective is Duolingo in developing the general L2 proficiency and receptive and productive vocabulary knowledge of L1-Spanish learners of English at a basic proficiency level? How does it compare to the L2-English development of similar learners receiving face-to-face classroom instruction?

RQ2. How do learner-related factors such as total time spent studying the course and level of motivation associate with the L2 proficiency and lexical development of learners in the Duolingo and classroom-based courses?

Following Sudina and Plonsky's (2024) approach, this study follows a "natural experiment", an observational study which describes "any event not under the control of a researcher that divides a population into exposed and unexposed groups" (Craig et al, 2017, p.2). Researchers exert less control over the intervention and use this natural variation to examine the impact of an event on the target outcome. This method enhances ecological validity compared to lab-based L2 instruction studies (Rogers & Cheung, 2021) and is well-suited for comparing app-based and classroom-based learning where researcher control is limited (Loewen et al., 2020). This study qualifies as a natural experiment because it follows a pretest-posttest design, takes place in authentic, less controlled settings, and involves participants who self-selected their mode of instruction (Craig et al., 2017).

## Methodology

### Participants

The participants were adult ( $M = 45.4$  years) Spanish speakers learning L2 English at a basic proficiency level (A2 CEFR) under two self-selected instruction modes: Duolingo or face-to-face, classroom-based instruction. Duolingo participants were invited to take part in the study when they were at the end of level A1.2 and beginning A2.1 (Units 45-46) of the English-for-Spanish course, and thus had previous experience with the app. Eligible participants were adults residing in non-English-speaking European countries, self-assessed as having basic English proficiency, and using Duolingo exclusively for English learning. They committed to 30 minutes of daily English study on the platform (3-4 hours weekly) for 16 weeks. Upon completion, participants received 100 euros in compensation.

Classroom participants were recruited from the Official Language School in Spain, a public institution regulated and subsidized by the Spanish Ministry of Education. Unlike private institutions, the school offers affordable, extra-curricular language courses to adults and awards official language certificates. As per the institution's policy, enrolment fees are non-refundable, which may influence students' decisions to complete a course, especially as compared to app-based learning. Participants were beginning the A2-level course, receiving 4 hours of weekly face-to-face instruction. They reported receiving only classroom-based English learning and committed to regular attendance. The institution received 700 euros for English learning materials in exchange for their collaboration.

After initial screening, 544 participants (188 classroom and 356 Duolingo learners) were invited to the pretest (see *Procedure*). However, 207 participants (72 classroom and 135 Duolingo learners) dropped out during the study period (i.e., ceased consistently attending the lessons [ $n=58$ ] or using the app [ $n=132$ ]) and/or did not complete the posttest ( $n=14$  Classroom and  $n=3$  Duolingo), resulting in a final participant pool of 337 learners. Only participants who completed both the pretest and posttest were included in the analyses. Participants in the Duolingo ( $k = 221$ ) and classroom ( $k = 116$ ) groups were matched in age ( $M = 45.5$ ,  $U = 12570.0$ ,  $z = -.292$ ,  $p = .770$ ), and both included a greater percentage of female than male learners (52.9% and 66.4% in each group, respectively). Both groups reported travelling and leisure as the main reasons for studying English, followed by job-related purposes. All participants lived in Spain, except for 4 Duolingo students that lived in Germany, Italy, the Netherlands and Serbia. See [Appendix A](#) (Table 1) in the [Supplementary Materials](#) online for detailed participant data and demographics by group.

### Instruments

This study employed standardized measures of English proficiency and vocabulary knowledge to establish generalizations across studies (Jiang, Rollinson, Plonsky, et al., 2021; Rachels & Rockinson-Szapkiw, 2018).

### Oxford Placement Test (OPT)

The OPT (Allen, 2004) is a standardized measure of L2-English ability. It comprises two sections: (1) *Listening*, which assesses students' general listening ability by choosing the correct word heard in short sentences; and (2) *Grammar*, which measures students' grammatical knowledge in context via items that require reading short sentences and choosing right answer. These tasks (Figure 2) are similar to the exercises that learners of lower proficiency levels are familiar with (García Botero et al., 2019). The test takes approximately 60 minutes to complete (~13 mins for the Listening part, and 50 mins for the Grammar part), and each section is scored over 100 points (1-0 per item) to produce a total aggregate score out of 200. The scores align with CEFR from pre-A1 to C2, accurately distinguishing between levels at the lowest proficiency, making it ideal for this study. The test was validated over 5 years with multilevel samples of students from more than 40 nationalities, and the results calibrated onto the CEFR (Allen, 2004). Independent studies also confirm its reliability ( $\alpha=.809$ , Wistner et al., 2009), making it a reliable instrument for examining L2-English proficiency (Borràs & Llanes, 2020). Internal consistency reliability was high for both groups (Duolingo's Cronbach's alpha ( $\alpha$ ) = .81, Classroom's  $\alpha$  = .82).

#### Figure 2

*OPT Sample Items (Listening and Grammar, Respectively)*

1. Water **is to boil / is boiling / boils** at a temperature of 100°C.

1. I gather you've been having trouble with your **earring / hearing**

### Vocabulary Levels Tests

Two untimed standardized tests were used to assess vocabulary development. The Updated Vocabulary Levels Test (uVLT; Webb et al., 2017) assesses receptive vocabulary knowledge (meaning recognition) of the most frequent 5,000 words (1,000-5,000 frequency levels), making it ideal for lower-level English learners. It uses a word-matching format to minimise guessing (Kremmel, 2020) (Figure 3). Given participants' low proficiency, only the 1,000–3,000 frequency levels were administered, comprising 90 items (30 per section) that were score dichotomously. Maximum test score was 90 (Duolingo  $\alpha$  = .84, Classroom  $\alpha$  = .81).

#### Figure 3

*uVLT Sample Items*

	game	island	mouth	movie	song	yard
land with water all around it		✓				
part of your body used for eating and talking			✓			
piece of music					✓	

The Productive Vocabulary Levels Test (PVL; Laufer & Nation, 1999) measures controlled productive knowledge (ability to recall the L2 forms) of the most frequent 2,000, 3,000, 5,000 and 10,000 words, plus academic vocabulary (each in a different section). Each of the five sections contains 18 items, where learners fill-in gapped English sentences with the correct word, based on the initial letters and context provided (Figure 4). Given the focus on lower-proficiency learners, only the 2,000 and 3,000-word levels were used (36 items total). Responses were scored dichotomously, and minor errors including misspellings (e.g., *apartament* for *apartment*) or grammatical infelicities (e.g., *\*this skirts*) were accepted if the intended word was clear. Responses were double-scored for inter-rater reliability, with an agreement of 90%. The maximum score in this test was 36 (Duolingo  $\alpha$  = .82, Classroom  $\alpha$  = .80).

## Figure 4

### *PVLT Sample Item*

This year long *sk*\_\_\_\_\_ are fashionable again.

### **Background and Motivation Questionnaire**

Following previous app-based research (Loewen et al., 2020; Jiang et al., 2021), a researcher-developed survey collected data on participants' language background, self-rated proficiency, prior English learning, and reasons for learning it. Additionally, questions regarding their level of engagement with the lessons and motivation for language learning before and after the course were included. Post-test items included additional reflection questions on out-of-class L2 engagement, learning progress and level of enjoyment with the course (see Hwang et al., 2024 for a review of the limitations of self-reported measures in MALL). Questions followed a 7-point Likert scale (see [Appendix B](#)).

### **Procedure**

The study involved three phases:

#### **Screening and Pretesting**

Eligible participants provided informed consent and completed a pretesting session to assess prior English knowledge and motivation. This included a background and motivation questionnaires, proficiency test (OPT) and vocabulary tests (PVLT and uVLT). Duolingo learners completed the tests individually online via Qualtrics, with remote proctoring using Hubstaff, which captured screenshots to monitor learner activity and ensure test-score validity. Duolingo participants were informed in advance of technical requirements (a computer with reliable internet access and sound). Classroom participants completed the tests in person during their lessons, supervised by the researcher for ecological validity. All instructions were given in Spanish. The pretesting stage took ~1.5 hours to complete, and the researchers controlled completion time following test guidelines (i.e., 60 minutes for the OPT) and piloting information (15–20 minutes for the uVLT and PVLT). Participants were informed of expected durations before each task.

#### **Learning Period**

Following prior app-based studies (Loewen et al., 2020; Lord, 2015; Rachels & Rockinson-Szapkiw, 2018), participants engaged in their respective language courses (Duolingo or classroom) over 16 weeks between pretesting and posttesting. To ensure comparable time spent in the language course in both groups, Duolingo participants were instructed to study about 30 minutes per day, aiming for 3–4 hours weekly. This was tracked via self-reported engagement questions. Duolingo data analytics were also collected to offer objective usage data. However, due to technical issues with recording accuracy throughout the duration of the study, this data was not employed in analyses to ensure the reliability and validity of the findings. No minimum study hours were enforced, but 73.8% of learners studied at least 2–3 hours per week, and 53.7% exceeded 3–4 hours. Duolingo participants were instructed to progress along the course content as much as possible, instead of only reviewing known material (Jiang et al., 2021a), and the progression made on the course units as well as the learning gains indicate this was typically adhered to. Classroom participants were asked to attend most lessons and follow the teachers' study plans (~60 hours of instruction in total for half an academic semester). Teachers confirmed that the classroom participants included in the analysis attended the course regularly (>80%). During the 16-week study period, the researchers maintained frequent communications with the Duolingo learners and classroom teachers. On week 14, a reminder was sent that the posttest was due in week 16.

## Posttesting

After the learning stage, participants completed the same tasks and questionnaire as in the pretest in the same format and order. Additional post-testing questions were added to the questionnaire asking participants whether they made use of any other courses or programs than the target ones during the duration of the study, to reflect on their learning experience and self-assess their English progress. Posttesting was completed in approximately 1.5 hours. This stage concluded with the conferral of participant/institution compensation.

## Analyses

Statistical analyses were conducted in R (version 4.1.1, R Core Team, 2021). All variables were centered and normalized to aid interpretation and reduce collinearity. General linear models (GLMs)<sup>2</sup> were fitted using the *glm* function to evaluate the effect of Mode of Instruction (Duolingo vs. Classroom) and learner factors on L2 proficiency and lexical development in English (RQs 1-2). Separate models were fitted for each dependent variable (OPT, uVLT, and PVLТ). Learner-related factors were modelled as predictors to isolate and better understand the effect of Mode of Instruction on any learning gains. Model fitting started from a core model including fixed effects of potential explanatory variables: total time spent learning English per week, interest in learning English and in the lessons, and motivation for learning English and in the lessons. Pretest scores were included as covariates to control for initial differences. We followed a step-by-step backwards model selection procedure using model comparison based on likelihood ratio tests, iteratively removing non-significant predictors to streamline the model and enhance its interpretability (Plonsky & Ghanbar, 2018), until reaching the final model with the lowest AIC value. Model fit was assessed using R<sup>2</sup> values from the performance package (Lüdecke et al., 2021). Pairwise comparisons between groups were conducted using the emmeans function (Version 1.10.2, Lenth, 2024).

## Results

### L2 Proficiency and Lexical Development

Table 1 shows that, overall, the participants' mean scores across all language tests were higher in the Duolingo group than the Classroom group, both at pretest and posttest. However, there was an exception with the OPT Listening test, in which classroom learners outperformed Duolingo learners prior and after the study period. This might be explained by the classroom learners' higher viewing and listening exposure to English prior and after the treatment (see Appendix C1, Table 1).

Wilcoxon signed-rank tests (data non-normally distributed) were independently conducted for each group to assess gains between the pretest and posttest. The tests revealed statistically significant gains across all linguistic measures for both the classroom and Duolingo groups ( $p < .05$ , see Table 2 for exact  $p$  and  $d$  values). This indicates that participants' general L2 proficiency and vocabulary knowledge improved significantly during the learning period, although the effect size (calculated using Cliff's delta) was small ( $< .40$ , Plonsky & Oswald, 2014).

In both groups, learning gains were more evident in the two vocabulary measures than the OPT. For the classroom learners, the difference in uVLT scores between pretest and posttest was 8.5% (63.2–54.7) and for the PVLТ 7.5% (31.1–26.3). In raw figures, this means a learning gain of 7.7 words in the uVLT and 2.7 words in the PVLТ standardized measures. Since in the uVLT test 30 items represent 1,000 words in each frequency band, this translates to an average gain of 256.7 real words at the receptive level of mastery ( $7.7 \times 33.3$ ). In the PVLТ, each item in the test represents approximately 55.5 words in a frequency band, meaning that classroom learners gained approximately 150 real words on average ( $2.7 \times 55.5$ ) at the productive level of mastery.

**Table 1***Descriptive Statistics for Language Test Scores (N = 337)*

Test	Classroom ( <i>k</i> =116)			Duolingo ( <i>k</i> =221)		
	<i>M</i> (%)	<i>SD</i>	Range	<i>M</i> (%)	<i>SD</i>	Range
OPT Total						
Pretest	49.3	8.1	28.5–71	54.6	9.9	13–74.5
Posttest	53	9	14.5–77	58.5	8.9	32.5–77.5
OPT Listening						
Pretest	63.8	8.2	42–82	63	12.2	10–85
Posttest	67.2	6.6	47–87	65	10.4	30–84
OPT Grammar						
Pretest	34.8	11	8–66	46.3	11.6	12–75
Posttest	39.7	12.3	8–74	52.1	10.8	27–86
uVLT						
Pretest	54.7	18.4	8.9–92.2	72.6	15.1	34.4–100
Posttest	63.2	17.5	17.8–93.3	80.2	13.6	40–100
PVLT						
Pretest	23.6	11.5	0–61	33.2	13.5	5.6–83.3
Posttest	31.1	10.6	11.1–58.3	39.9	15.5	8.3–88.9

**Table 2***Significance and Effect Size for the Paired Samples (Pre-Post) Test Contrasts*

Group	OPT Total		uVLT		PVLT	
	<i>p</i>	<i>d</i>	<i>p</i>	<i>d</i>	<i>p</i>	<i>d</i>
Classroom	<.001	0.30	<.001	0.31	<.001	0.18
Duolingo	<.001	0.33	<.001	0.46	<.001	0.23

The picture is similar for the Duolingo learners, with a difference of 7.6% in the uVLT (80.2–72.6) and 6.7% in the PVLT (39.9–33.2) from pretest to posttest. In raw figures, this is an average gain of 7 words in the uVLT task and 2 in the PVLT task, translating into 233.3 real words gained at receptive knowledge and 111.1 at productive level.

## L2 Motivation and Engagement

Table 3 shows that participants' motivation and interest in the L2 and target course were very high in both groups before and after the instruction period. However, both groups experienced a decline in motivation and interest after the study period.

**Table 3***Interest and Motivation Prior and After Instruction*

Characteristic	Classroom ( <i>k</i> =116)			Duolingo ( <i>k</i> =221)		
	<i>M</i>	<i>SD</i>	Range	<i>M</i>	<i>SD</i>	Range
Interested in English						
pretest	6.69	.55	4–7	6.81	.51	2–7
posttest	6.65	.59	4–7	6.75	.49	5–7
Motivated in English						
pretest	6.61	.68	4–7	6.62	.66	2–7
posttest	6.53	.71	4–7	6.47	.84	2–7
Interested in the lessons						
pretest	6.72	.49	5–7	6.71	.62	1–7
posttest	6.54	.73	4–7	6.60	.62	4–7
Motivated in the lessons						
pretest	6.69	.57	4–7	6.70	.60	1–7
posttest	6.43	.91	2–7	6.57	.75	2–7

*Note.* Rated on a 1-7 Likert-Scale (1= Strongly disagree, 7 = Strongly agree).

A two-way repeated measures ANOVA was run to estimate changes in motivation from pretest to posttest, examining sustained motivation (within-subject), group (between-subject), and their interaction. Results showed a significant main effect of posttest motivation within-subjects ( $p < .001$ ) but no significant effect of group ( $p = 0.486$ ) or interaction effect ( $p = 0.149$ ). The effect size ( $\eta^2 = 0.26$ ) indicated a moderate effect (Plonsky & Oswald, 2014). This suggests participants' motivation significantly decreased from pretest to posttest on an individual level, not a group level.

A Welch Two Sample *t*-test found no significant differences in interest and motivation between Classroom and Duolingo participants before the study ( $p = 0.060$  and  $0.965$ , respectively). However, a significant difference was found ( $p < .001$ ) between the groups when comparing posttest motivation (i.e., sustained motivation). The median score for the Classroom group was 5 (Q1 = 4, Q3 = 6), and for the Duolingo group it was 6 (Q1 = 5, Q3 = 7), indicating higher average posttest motivation among Duolingo learners (medium effect size,  $r = 0.49$ ). Thus, both groups initially had similar motivation levels, but Duolingo learners had slightly higher sustained motivation levels.

To assess course engagement, participants were asked to self-report their average weekly study time. Table 4 shows classroom learners averaged 4-5 hours per week (including lesson time), while Duolingo learners spent an average of 2.5 hours. A Welch Two-Sample *t*-test indicated a significant difference ( $p < .001$ ), with classroom learners studying more. This suggests the Duolingo group, on average, did not meet the 3–4-hour weekly goal, although 53.7% of learners reported studying 3 or more hours per week, meeting the study target. In contrast, classroom participants attended at least 80% of the lessons (~48 hours of the expected 60), suggesting a greater engagement with the course.

**Table 4***Average Time Studying English Per Week in the Target Mode*

Characteristic	Classroom ( <i>k</i> =116)			Duolingo ( <i>k</i> =221)		
	<i>M</i>	<i>SD</i>	Range	<i>M</i>	<i>SD</i>	Range
Average time studying English per week <sup>a</sup>	6.17	1.6	3–8	4.37	1.4	1–6

Note. <sup>a</sup>Values represent: 1 = 1–30 mins, 2 = 31–60 mins, 3 = 1–2h, 4 = 2–3h, 5 = 3–4h, 6 = 4–5h, 7 = 5–6h, 8 = 6 + h

### Effect of Mode of Instruction and Learner Factors on L2 Development

To determine comparability between the Duolingo and Classroom groups, Mann-Whitney *U* pairwise comparisons were conducted on all pretest scores (OPT, uVLT, and PVLТ) (preliminary analyses in [Appendix C2](#)). Results showed a significant advantage for the Duolingo group across all linguistic measures ( $p < .001$ ). Despite both groups starting an A2-level course, the Duolingo participants' level seemed, on average, higher than estimated (mean OPT score of 109.3, corresponding to A2-level [105–119 OPT scores = A2 CEFR]). This suggests that Duolingo learners might have been at early stages of the A2-level at pretest, as compared to the classroom learners who had just achieved the A1 level (OPT mean scores was 98.6 [90–104 = A1]). This finding reaffirmed the decision to include pretest scores as covariates in the GLMs to control for pretest differences<sup>3</sup>. What follows are the model results by outcome measure, with the models detailed in [Appendix C3](#).

#### OPT

A GLM was calculated with the OPT total posttest scores as the dependent variable. The model revealed a main effect for mode of instruction ( $\beta = 0.18$ ,  $SE = 0.08$ ,  $p = .023$ ), with the Duolingo group scoring higher than the Classroom group. Pretest scores were also a significant predictor of learning ( $\beta = 0.71$ ,  $SE = 0.04$ ,  $p < .001$ ). Self-reported interest and motivation had also a significant effect on OPT posttest scores ( $\beta = 0.11$ ,  $SE = 0.05$ ,  $p = .026$  and  $\beta = -0.10$ ,  $SE = 0.05$ ,  $p = .044$ , respectively). Time spent learning per week did not predict learning, suggesting an effective control over study time for both groups. The model explained 56% of the variance (adjusted- $R^2 = .56$ ).

GLMs were also fitted for each OPT section. The Grammar model showed significant effects of mode of instruction ( $\beta = 0.29$ ,  $SE = .07$ ,  $p < .001$ ), pretest scores ( $\beta = 0.73$ ,  $SE = 0.04$ ,  $p < .001$ ), self-reported interest in learning English ( $\beta = 0.11$ ,  $SE = 0.04$ ,  $p = .014$ ), and motivation ( $\beta = 0.15$ ,  $SE = 0.04$ ,  $p < .001$ ). The Duolingo group displayed a significantly greater improvement in grammar than the classroom learners (adjusted- $R^2 = .67$ , explaining 67% of variance in the model). The Listening model revealed significant effects of pretest Listening ( $\beta = 0.53$ ,  $SE = 0.05$ ,  $p < .001$ ) and mode of instruction ( $\beta = -0.20$ ,  $SE = 0.10$ ,  $p = .036$ ). In this case, however, the Classroom group outperformed the Duolingo learners, suggesting that classroom learners exhibited a greater improvement in their listening skills compared to the Duolingo learners. The model explained 39% of the variance (adjusted- $R^2 = .39$ ).

#### Vocabulary Measures

A GLM fitted for the uVLT posttest scores revealed that mode of instruction predicted learning, favoring the Duolingo group ( $\beta = 0.24$ ,  $SE = 0.07$ ,  $p < .001$ ). Performance at pretest was also significant ( $\beta = 0.66$ ,  $SE = 0.05$ ,  $p < .001$ ), indicating higher initial receptive vocabulary scores associated with higher posttest scores regardless of group. Given the relationship between receptive and productive vocabulary knowledge (González-Fernández, 2025), PVLТ pretest scores were included as a covariate.

The effect of pretest PVLT scores was also significant ( $\beta = 0.16$ ,  $SE = 0.05$ ,  $p < .001$ ). The model reached 73% of the variance (adjusted- $R^2 = .73$ ). In the GLM calculated for PVLT posttest scores, mode of instruction did not have a significant effect on participants' performance ( $\beta = 0.04$ ,  $SE = 0.09$ ,  $p = .636$ ), unlike in previous models. However, PVLT pretest scores significantly predicted posttest outcomes ( $\beta = 0.61$ ,  $SE = 0.06$ ,  $p < .001$ ), and uVLT pretest scores also had a significant effect ( $\beta = 0.14$ ,  $SE = 0.06$ ,  $p = .030$ ). The model explained 53% of the variance in productive vocabulary outcomes (adjusted- $R^2 = .53$ ).

Post hoc comparisons using estimated marginal means were conducted to assess group differences. The results revealed significant differences between the two groups (see [Appendix C](#), Table 6). Specifically, Classroom learners exhibited significantly lower scores compared to Duolingo learners in OPT total ( $p = .028$ ), OPT grammar ( $p < .001$ ), and uVLT ( $p = .006$ ). Conversely, Classroom learners achieved significantly higher scores than Duolingo learners in OPT listening ( $p = .036$ ). However, differences in PVLT scores between the groups did not reach statistical significance ( $p = .635$ ).

## Discussion

This study compared motivation and L2 development in Spanish-speaking beginner learners of English via classroom-based or Duolingo instruction, controlling for program length and study time. Key findings and implications are discussed below.

### RQ1: Effectiveness of Duolingo for L2 Development

There are two main findings connected to this RQ. First, both modes of instruction (i.e., classroom-based vs. app-based) led to significant improvement in L2 proficiency and lexical knowledge (as measured by standardized tests) after 16 weeks of studying. This finding corroborates the beneficial and crucial effect of deliberate instruction in L2 development, regardless of type (Norris & Ortega, 2000; Rachels & Rockinson-Szapkiw, 2018). Students showed gains on vocabulary knowledge, both at receptive and productive mastery, corroborating prior findings that MALL apps are effective for vocabulary development (Shortt et al., 2023). As expected, improvement was greater in receptive than productive vocabulary in both modes of instruction (González-Fernández, 2025). In a previous app-based study, Jiang et al., (2021b) found that productive vocabulary exhibited the weakest improvement among all linguistic measures. All this suggests that while MAAL apps facilitate overall lexical development, a greater focus on productive vocabulary knowledge in these apps would be beneficial (Lin & Lin, 2019).

Secondly, after accounting for pretest differences in L2 knowledge between the groups, the results showed a significant main effect of Mode of Instruction on learning gains across all measures, except the PVLT. A general advantage of Duolingo learners over classroom learners was found on OPT Total, OPT grammar, and uVLT after the 16-week learning period. However, classroom learners significantly outperformed Duolingo students on the OPT Listening task after the study period. Interestingly, although Duolingo learners did not experience significant gains in the listening task, 60% self-reported listening as the most developed skill through instruction (see [Appendix C1](#)). This discrepancy highlights a potential limitation of listening activities in app-based learning and suggests that listening gains in such environments may be overestimated by L2 learners. This finding corroborates that L2 development is significantly affected by type of instruction (Norris & Ortega, 2000).

Importantly, the findings suggest a potential advantage of using Duolingo for developing certain aspects of L2 proficiency, particularly receptive grammar and vocabulary knowledge, at least for low-proficiency learners. This aligns with prior research showing a beneficial impact of MALL apps, specifically Duolingo, on various language competencies (see Shortt et al.' 2023 systematic review). The results indicate that when study time is comparable in both modes of instruction app-based instruction can outperform classroom instruction on certain linguistic aspects as measured by the target written language tests.

While our models controlled for pretest differences between the groups, it is important to note that Duolingo learners may have had greater initial L2 exposure. This may reflect differences in course structure and timing: whereas the classroom course resumed after a summer break, potentially limiting recent L2 exposure, the Duolingo course progressed from one level to the next without interruption. This difference in initial input may have contributed to Duolingo learners' performance gains.

This study also found that both instruction modes were similarly effective in developing productive vocabulary knowledge, which is key for writing and speaking. The PVLТ is strongly associated with L2 production proficiency (Suzuki & Kormos, 2023), suggesting that both modes of instruction have similar potential to promote productive skills at a basic proficiency level. However, further research is needed to directly assess learners' productive skills to test this hypothesis. Additionally, Spanish and English share many cognate words (words that derive from the same original language). Around 34–37% of English words have Spanish cognates (Lubliner & Hiebert, 2011), which may have supported the strong vocabulary gains observed in this sample, especially in productive knowledge.

Notably, classroom instruction was more beneficial for the development of listening skills than app-based instruction, likely due to increased opportunities for spoken interaction and oral practice in classroom settings. This aligns with Lord (2015), who found that app-based learners struggled more with conversation than classroom learners. Similarly, participants in Loewen et al. (2020) perceived app-based learning to be effective for grammar and vocabulary development, but less so for speaking skills. One advantage of apps is their flexibility, as they are often used during “dead time” such as commuting (Wu, 2015). However, when using apps in this manner, learners might skip or mute audio tasks, reducing listening exposure<sup>4</sup>. Apps could further enhance and control the listening exposure students receive by including self-evaluation exercises to track their audio usage or adding progression locks requiring completion of a minimum number of listening-only tasks before advancing to the next lesson, while maintaining enough flexibility to prevent frustration and disengagement. In contrast, given the communicative approach of the classroom lessons, it is likely that classroom learners received more extensive and less-controlled exposure to spoken English than the Duolingo learners, which primarily engaged in controlled, scripted listening exercises. Additionally, classroom learners reported higher weekly out-of-class exposure to oral English through videos, TV, radio, podcasts, and music than Duolingo learners. These differences may explain classroom learners' greater listening development. Previous research has shown that regular out-of-class listening exposure can effectively develop listening skills (e.g., Muñoz & Cadierno, 2021). These findings also suggest that each instructional mode may suit different learner goals. Apps may be especially effective for developing foundational linguistic competence, specifically receptive knowledge of vocabulary and grammar, either for independent learners with limited access to formal education or as a complement to classroom-based training. On the other hand, teacher-led classroom instruction may better support the development of communicative competence through more interactive, social and interpersonal language use.

## **RQ2: L2 Motivation and Engagement**

The impact of motivation and engagement on L2 development across both instruction modes showed mixed results. Classroom learners reported spending more time studying English per week than Duolingo learners, and their course attendance was >80%, indicating a greater level of engagement with the course potentially fostered by as a sense of commitment to the teacher and peers. Yet, 54% of Duolingo learners still met the study target of 3–4 hours per week. This contrasts with prior app-based research, which showed low engagement and persistence with the app course (e.g., only 22% of learners met the study target in Loewen et al. 2019). This might explain why weekly study time did not predict learning gains, despite its known influence on L2 development (García Botero et al., 2019; Loewen et al., 2019, 2020).

Interestingly, despite spending more time studying English, classroom learners showed somewhat less L2 improvement than Duolingo learners. One explanation is that classroom learners might not be as engaged with each individual in-class activity. In contrast, app-based learners must actively engage with tasks to some extent, even if only for short periods, to complete lessons and progress on the course. Another factor

might be the individualized feedback after each task in app-based learning, which provides learners with a sense of achievement and progression after each session. In a classroom setting, individual feedback may not be possible for every task. Thus, although Duolingo learners spent less time studying on average, their study time may have been more effective. Importantly, the two instructional settings are conceptually and qualitatively different, and learner experience likely varied in each mode despite controlling for study time. Thus, differences in outcomes should be interpreted as reflecting structural and contextual variation in each instruction mode, rather than as one mode being inherently superior.

Concerning motivation, the study shows that learners in both groups had high levels of motivation in the lessons and the language before the learning period, but this motivation decreased significantly afterward. Due to its dynamic nature, it is common for motivation levels to change over time (Dörnyei, 2019), and this study shows that this occurs in both app-based and classroom-based instruction. The findings indicate that Duolingo learners showed slightly higher motivation levels at posttest compared to classroom learners. This contrasts with previous studies reporting low long-term motivation in app-users (e.g., García Botero et al., 2019; Loewen et al., 2019). Gamification elements in Duolingo may have made the learning experience more appealing and enjoyable for this group (James & Mayer, 2019). Additionally, as suggested by He and Loewen (2022), explicit weekly goal-setting for app-based learners in this study (i.e., specifying and recording a target study time), might have contributed to the increased motivation and engagement. Overall, the current findings highlight the motivational benefits of Duolingo for L2 learning. However, this finding may be partially explained by the higher attrition in the Duolingo group ( $n = 135$ ) compared to the classroom learners ( $n=72$ ), which may have led to only the most motivated app-based learners remaining in the study. It is also possible that the participants' choice of each method of instruction may reflect underlying qualitative differences, such as long-term goals or need for an official language accreditation, which may have influenced their motivation and engagement in the course beyond what our data captured. For example, 25% of classroom participants reported studying English for educational purposes compared to 12% of Duolingo learners (see [Appendix A](#), Table 1). Future research could further explore what motivates the choice of instruction and how these factors shape motivation, engagement and language development.

Finally, regarding the influence of motivation on linguistic achievement, self-reported motivation in this study only influenced the OPT Total, OPT Grammar and PVLT scores at posttest. This contrasts with claims that motivation strongly predicts L2 performance (Loewen, 2020). This finding may reflect a discrepancy between reported motivation and actual behavior. For example, García Botero et al. (2019) found that while students reported high motivation and positive attitudes toward using apps, when interviewed they showed mixed perceptions and a lack of interest using the app long term when interviewed. Future research could explore this further by collecting interview data to better understand participants' motivation. Another limitation is that Duolingo currently only offers US-accented English instruction. This restricts learners' exposure to regional accent variation, which may have influenced their motivation, attitudes or expectations. Future research should explore how this one-accent option affects engagement and listening development in app-based learning. In addition, this study relies on self-reported measures of language exposure, engagement, and motivation, which may be subject to retrospection bias (recalling past engagement more positively than it was) or overestimation (see Hwang et al., 2024). Future research should incorporate objective measures to better capture actual behavior and engagement.

## Conclusion and Future Directions

The study shows that Duolingo holds promise as an alternative and complementary means of access to L2 learning for developing general written English proficiency and vocabulary in learners at a basic proficiency level. Given Duolingo's popularity, the findings offer broader insights into the effectiveness of app-based versus classroom-based instruction for L2 acquisition. Yet, further research is needed to advance our understanding in this area. The results reflect short-term gains at one post-intervention

testing point. Future research should assess the sustainability of these gains longitudinally over a more extended period and with delayed post testing. Despite our efforts to align the beginning of the Duolingo and classroom courses, higher pretest scores in the Duolingo group suggest greater prior exposure to English, potentially influencing their learning trajectory. Future studies could use stratified matching to reduce initial differences and better isolate instructional effects. The motivation survey used in this study adopted a broad conceptualization of motivation and was limited to participants who completed the course, likely those with higher sustained motivation compared to dropouts. Future studies should employ validated measures of motivation and engagement and aim to minimize learner attrition to better understand their role in app-based learning. Since the study focused on L1-Spanish EFL learners, it is unclear whether the results generalize to learners from non-cognate language backgrounds, such as Chinese or Arabic. Cognateness is known to facilitate L2 acquisition, especially vocabulary (e.g., González-Fernández, 2025). Future research should explore how this factor influences learning across app-based and classroom instructional modes. It should also assess real-life communication skills, such as speaking, to fully capture L2 proficiency development beyond L2 grammar and lexical development.

Taken together, these findings highlight the complementary strengths of different instructional modes. While apps can support personalized and adaptative vocabulary and grammar learning, classroom settings appear better suited to developing listening and communicative skills. These results support the combination of mobile-assisted learning and classroom-based instruction for L2 development, with practical implications for research, learning and teaching.

## Acknowledgements

This project was funded by a Duolingo Efficacy Study grant, but was planned, designed and conducted independently by the authors. We would like to thank Duolingo, and specially Dr. Xiangying Jiang and members of the Learning Science Lab, for their interest in the study and their support during its completion. We are also grateful to the three anonymous reviewers and the handling Editor for their insightful comments and suggestions on earlier versions of this manuscript. Finally, we are incredibly grateful to all the learners who saw the value of the study for their linguistics development and generously contributed their time, energy and knowledge, and to the teachers who helped us advertise and promote the study. Without their collaboration, this project would not have been possible.

## Notes

1. In this study, classroom-based instruction refers to face-to-face, teacher-led language learning that takes place within a formal educational setting, typically following a structured curriculum and schedule.
2. Models were fitted with a Gaussian distribution. Predictor and outcome variables were z-scored using the `scale()` function to improve interpretability and model convergence.
3. To explore potential lurking variables, we compared participants' educational background across groups as a proxy for socioeconomic status. No significant differences were found ( $\chi^2(2)=20.04, p = 0.078$ ), suggesting that educational background was comparable across instructional modes.
4. This interpretation draws on anecdotal evidence from some Duolingo users. Future research should explore app learners' behavior and engagement with listening tasks, possibly using stimulated recall.

## References

- Allen, D. (2004). *Oxford placement test*. Oxford University Press.
- Borràs, J., & Llanes, À. (2020). L2 reading and vocabulary development after a short Study Abroad experience. *Vigo International Journal of Applied Linguistics*, 17, 35–55. <https://doi.org/10.35869/VIAL.V0I17.1464>
- 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>
- Council of Europe. (2017). Common European framework of reference for languages: Learning, teaching, assessment—Companion volume with new descriptors. Retrieved from <https://rm.coe.int/common-european-framework-of-reference-for-languages-learningteaching/168074a4e2>
- Craig, P., Katikireddi, S. V., Leyland, A., & Popham, F. (2017). Natural experiments: An overview of methods, approaches, and contributions to public health intervention research. *Annual Review of Public Health*, 38, 39–56. <https://doi.org/10.1146/annurev-publhealth-031816-044327>
- Dehgan-zadeh, H., & Dehgan-zadeh, H. (2020). Investigating effects of digital gamification-based language learning: A systematic review. *Journal of English Language Teaching and Learning*, 12(25), 53–93. <https://doi.org/10.22034/ELT.2020.10676>
- Dörnyei, Z. (2019). Towards a better understanding of the L2 learning experience, the Cinderella of the L2 motivational self system. *Studies in Second Language Learning and Teaching*, 9(1), 19–30. <https://doi.org/10.14746/ssllt.2019.9.1.2>
- García Botero, G., Questier, F., & Zhu, C. (2019). Self-directed language learning in a mobile-assisted, out-of-class context: Do students walk the talk? *Computer Assisted Language Learning*, 32(1–2), 71–97. <https://doi.org/10.1080/09588221.2018.1485707>
- González-Fernández, B. (2025). How is vocabulary learnt? An acquisitional sequence of L2 word knowledge. *TESOL Quarterly*, 59(2), 755–784. <https://doi.org/10.1002/tesq.3366>
- Guaqueta, C. A., & Castro-Garces, A. Y. (2018). The use of language learning apps as a didactic tool for EFL vocabulary building. *English Language Teaching*, 11(2), 61–71. <https://doi.org/10.5539/elt.v11n2p61>
- Hwang, H. B., Coss, M. D., Loewen, S., & Tagarelli, K. M. (2024). Acceptance and engagement patterns of mobile-assisted language learning among non-conventional adult L2 learners: A survival analysis. *Studies in Second Language Acquisition*, 1–27. <https://doi.org/10.1017/S0272263124000354>
- He, X., & Loewen, S. (2022). Stimulating learner engagement in app-based L2 vocabulary self-study: Goals and feedback for effective L2 pedagogy. *System*, 105, Article 102719. <https://doi.org/10.1016/j.system.2021.102719>
- James, K. K., & Mayer, R. E. (2019). Learning a second language by playing a game. *Applied Cognitive Psychology*, 33(4), 669–674. <https://doi.org/10.1002/acp.3492>
- 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), 249–272. <https://doi.org/10.1558/cj.26704>
- Jiang, X., Rollinson, J., Chen, H., Reuveni, B., Gustafson, E., Plonsky, L., & Pajak, B. (2021a). How well does Duolingo teach speaking skills? *Duolingo Research Report DRR-21-02*.

- Jiang, X., Rollinson, J., Plonsky, L., Gustafson, E., & Pajak, B. (2021b). 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>
- Kremmel, B. (2020). Measuring vocabulary learning progress. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 406–418). Routledge.
- Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, 16(1), 33–51. <https://doi.org/10.1177/026553229901600103>
- Lenth, R. (2024). *emmeans: Estimated Marginal Means, aka Least-Squares Means* (version 1.10.2). Retrieved from <https://CRAN.R-project.org/package=emmeans>
- Lin, J.-J., & Lin, H. (2019). Mobile-assisted ESL/EFL vocabulary learning: A systematic review and meta-analysis. *Computer Assisted Language Learning*, 32(8), 878–919. <https://doi.org/10.1080/09588221.2018.1541359>
- Loewen, S. (2020). *Introduction to instructed second language acquisition* (2<sup>nd</sup> ed.). Routledge.
- 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](https://doi.org/10.1111/modl.12234_3)
- Lubliner, S. & Hiebert, E. H. (2011) An analysis of English–Spanish cognates as a source of general academic language. *Bilingual Research Journal* 34, 76–93. <https://doi.org/10.1080/15235882.2011.568589>
- Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, 6(60), 3139. <https://doi.org/10.21105/joss.03139>
- Muñoz, C., & Cadierno, T. (2021). How do differences in exposure affect English language learning? A comparison of teenagers in two learning environments. *Studies in Second Language Learning and Teaching*, 11(2), 185–212. <https://doi.org/10.14746/ssllt.2021.11.2.2>
- Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, 50(3), 417–528. <https://doi.org/10.1111/0023-8333.00136>
- Plonsky, L., & Ghanbar, H. (2018). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R2 values. *The Modern Language Journal*, 102(4), 713–731. <https://doi.org/10.1111/modl.12509>
- 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. (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. <https://doi.org/10.1080/09588221.2017.1382536>

- Rogers, J., & Cheung, A. (2021). Does it matter when you review?: Input spacing, ecological validity, and the learning of L2 vocabulary. *Studies in Second Language Acquisition*, 43(5), 1138–1156. <https://doi.org/10.1017/S0272263120000236>
- Saito, K., Dewaele, J., Abe, M., & In'nami, Y. (2018). Motivation, emotion, learning experience, and second language comprehensibility development in classroom settings: A cross-sectional and longitudinal study. *Language Learning*, 68(3), 709–743. <https://doi.org/10.1111/lang.12297>
- Suzuki, S., & Kormos, J. (2023). The multidimensionality of second language oral fluency: Interfacing cognitive fluency and utterance fluency. *Studies in Second Language Acquisition*, 45(1), 38–64. <https://doi.org/10.1017/S0272263121000899>
- 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>
- 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>
- Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test. *ITL-International Journal of Applied Linguistics*, 168(1), 33–69. <https://doi.org/10.1075/itl.168.1.02web>
- Wistner, B., Sakai, H., & Abe, M. (2009). An analysis of the Oxford Placement Test and the Michigan English Placement Test as L2 proficiency tests. *Bulletin of the Faculty of Letters, Hosei University*, 58(2), 33–44. <https://doi.org/10.15002/00004453>
- Wu, Q. (2015). Pulling mobile assisted language learning (MALL) into the mainstream: MALL in broad practice. *PloS one*, 10(5), e0128762. <https://doi.org/10.1371/journal.pone.0128762>

## Appendices

Find appendices in the [Supplementary Materials](#) online.

## About the Authors

Beatriz González-Fernández is an Associate Professor in Applied Linguistics and TESOL at the University of Sheffield, UK. Her research interests include the learning and teaching of second/foreign languages, particularly vocabulary. She has published in relevant journals in the field, including *Applied Linguistics*, *TESOL Quarterly*, and *Studies in Second Language Acquisition*. Beatriz González-Fernández is the corresponding author.

**E-mail:** [b.gonzalez-fernandez@sheffield.ac.uk](mailto:b.gonzalez-fernandez@sheffield.ac.uk)

**ORCID:** <https://orcid.org/0000-0002-4370-1822>

Inés de la Viña is a Teaching Associate in Applied Linguistics at the University of Nottingham, specializing in bilingualism, cognitive development, and second language acquisition. She holds a PhD in Linguistics from the University of Kent and has extensive international teaching experience, integrating research-led, inclusive pedagogy with an active research profile in applied linguistics.

**ORCID:** <https://orcid.org/0000-0001-7019-5091>