Studies in Second Language Acquisition (2026), **48**, 680–707
doi:10.1017/S0272263126101521

RESEARCH ARTICLE

Comparing the effectiveness of Duolingo, Classroom instruction, and Classroom + Duolingo instruction conditions on beginner-level French language development

YouJin Kim[1], Caroline Payant[2], Stephen Skalicky[3] and Yoon Namkung[1]

[1]Applied Linguistics & ESL, Georgia State University, Atlanta, GA, USA; [2]Université du Québec à Montréal, Montréal, Canada and [3]Victoria University of Wellington, New Zealand
Corresponding author: YouJin Kim; Email: ykim39@gsu.edu

(Received 28 November 2024; Revised 10 November 2025; Accepted 02 December 2025)

Abstract

Over the past decade, there has been growing use of mobile-assisted language learning (MALL) applications such as Duolingo. The effectiveness of MALL applications is thus of great interest among language acquisition researchers and practitioners. This study compared French language development among beginners in three learning conditions: Classroom-Only (n = 58), Duolingo-Only (n = 65), and Classroom + Duolingo (n = 60). The Classroom-Only group completed a standard first-semester curriculum, the Classroom + Duolingo group used Duolingo as supplemental material, and the Duolingo-Only group learned exclusively through the app. All participants completed pretests and posttests measuring overall proficiency, grammar, vocabulary, pragmatics (*tu* vs. *vous*), and communicative competence over a 16-week period. Linear mixed-effects models revealed that all three groups showed significant improvement across nearly all measures between pretest and posttest, with similar magnitudes of improvement. The only exception was in learning *tu* vs. *vous* (pragmatic competence), where the Classroom + Duolingo group showed larger gains than Classroom-Only and Duolingo-Only groups. Results suggest that both traditional classroom instruction and Duolingo are comparably effective for beginning French language learners. Results are discussed in light of mobile app-based language learning with the potential role of learner characteristics.

Keywords: Beginner L2 French learners; Classroom instruction; Duolingo; Mobile-assisted language learning

Introduction

Research in second language acquisition (SLA) provides evidence that the use of mobile-assisted language learning (MALL) applications such as Duolingo, Busuu,

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives licence (http://creativecommons.org/licenses/by-nc-nd/4.0), which permits non-commercial re-use, distribution, and reproduction in any medium, provided that no alterations are made and the original article is properly cited. The written permission of Cambridge University Press or the rights holder(s) must be obtained prior to any commercial use and/or adaptation of the article.

https://doi.org/10.1017/S0272263126101521 Published online by Cambridge University Press

and Babbel is a beneficial method for language learning (Jiang et al., 2020; Rachels & Rockinson-Szapkiw, 2018; Vesselinov & Grego, 2012, 2016). These studies suggest that MALL application use affords an equal or greater amount of language learning opportunities for the development of reading and listening skills when compared to foreign language (FL) instruction in classroom contexts. However, the impact of MALL applications on the development of oral skills has received limited attention, partly due to the focus of MALL apps on receptive rather than productive skills (cf. Loewen et al., 2020; Kittridge et al., 2025a, 2025b). The recent development of Duolingo lessons targeting oral production in French provides a timely opportunity to assess the efficacy of MALL applications on productive skills and to compare the efficacy of these lessons to FL instruction alone and as supplemental instructional material in a classroom context. Accordingly, this study aims to compare the efficacy of Duolingo use, traditional classroom instruction in tandem with Duolingo use, and traditional classroom instruction alone on French language development among beginner-level learners. Because MALL applications are relatively affordable or offered at no cost, the efficacy of such educational resources is worth investigating for their potential to offer equitable language learning opportunities for people from diverse socio-economic backgrounds.

Language development using mobile apps

The use of MALL applications for language learning has been steadily increasing (Hwang et al., 2024; Loewen et al., 2019, 2020). Accordingly, research focusing on their use and efficacy for language learning has been on the rise in the domain of instructed second language acquisition (ISLA; Loewen et al., 2020; Smith et al., 2024; Sudina & Plonsky, 2024) (see Appendix A for a list of sample studies that have examined MALL apps and language learning). In our literature review, we included research on MALL applications that come with certain instructional curricula and did not consider studies focusing on certain technologies offered via MALL applications (e.g., voice recording applications). In this line of research, the length of study ranges from 8 to 16 weeks, which is the typical length of a regular semester for FL instruction at the university level. Previous research has mainly focused on beginner-level adult learners studying a limited range of target languages (e.g., French, Spanish, English, and Turkish), with a primary focus on receptive and grammatical competence (cf. Rachels & Rockinson-Szapkiw, 2018; Smith et al., 2024) and vocabulary knowledge (Enayat et al., 2025). In general, findings of the previous research have shown that language skills improve after using these mobile apps.

Notably, relatively few MALL application-based studies have focused on the development of oral skills. Of these studies, Loewen et al. (2020) examined the effectiveness of Babbel on communicative ability with English-speaking learners of Spanish. They found that the amount of usage time and learners' level of interest in learning Spanish were related to increases in oral proficiency based on the Oral Proficiency Interview Computer (OPIc) scores; however, evidence of oral skill development using ACTFL ratings was modest. Indeed, Loewen et al. (2020) reported that the outcomes for 12 participants went from having "no real functional ability to being able to communicate minimally by using a number of isolated words and memorized phrases" (p. 227). In other words, these modest gains did not prepare learners to hold conversations in the target language. Moreover, with such broad descriptors of language ability, it may be challenging to capture specific aspects of oral skill improvement, especially with low-level language learners.

https://doi.org/10.1017/S0272263126101521 Published online by Cambridge University Press

More recently, a Duolingo research report examined the efficacy of Duolingo on the development of speaking skills among Spanish and French language learners who completed beginner-level course materials (Jiang et al., 2021). Using the Pearson Versant Spanish and French Tests to assess speaking skills development, the findings showed that most participants achieved a level of A2 or above on the CEFR scale (Council of Europe, 2001) after using the app. As such, these studies suggest that MALL applications have a positive impact on the development of second language (L2) oral skills when using holistic measures from standardized tests. However, these tests fail to provide fine-grained information about the development of linguistic subsystems related to oral skills development.

Kittredge et al. (2024), in a Duolingo internal efficacy report, created customized oral tests for beginner learners of Spanish, French, and English. These tests placed learners in conversational contexts and recorded their production of different sentences in the target language. For example, one item informed test-takers that they were going to a party and prompted them to greet other attendees using Spanish. Oral responses were evaluated by trained raters using a two-point scale to assess communicative abilities. Kittredge et al. (2024) reported that after 4–6 weeks of Duolingo use, learners were able to initiate communication (basic greetings) and respond to prompting questions. However, these tests elicited responses to scripted items rather than natural conversation, limiting their ability to measure authentic conversational skills.

The instructional features of such language learning apps have been constantly upgraded as different educational technologies evolve, and the production language skills have been increasingly examined. For instance, Kittredge et al. (2025a) and Kittredge et al. (2025b) recently examined how generative artificial intelligence (AI) features within Duolingo lessons may influence learners' speaking proficiency. For instance, the Video Call feature provides real-time spoken conversation practice through auditory and written interaction with a character in real-life scenarios, giving GPT-generated feedback on accuracy and improvement tips. Findings revealed that over 1 month, low-intermediate learners of English using this feature demonstrated greater improvement in speaking proficiency compared to a control group using the free version of Duolingo without this AI feature (Kittredge et al., 2025b). Moreover, Kittredge et al. (2025a) examined beginner learners' self-efficacy (i.e., learners' belief that they can perform at a particular level on a task) after using Duolingo's Roleplay and Explain My Answer features. These AI features offer interactive speaking/writing practice through role-plays and personalized feedback on learner responses. The results showed that French and Spanish learners who used these AI features demonstrated significantly higher self-efficacy and reported feeling more prepared to use the target languages in real-life situations outside the app. While these studies are essential to understand how generative AI can further support MALL, the duration of Duolingo usage was relatively short (i.e., 1 month), limiting conclusions that can be made for long-term effectiveness of app usage and L2 development. Furthermore, speaking proficiency was measured using holistic speaking scores; thus it is difficult to determine which aspects of linguistic competence in learner speaking production improved.

Research has also examined whether certain language skills are more amenable to development than others with MALL applications (see Appendix A for a summary table of previous research). For example, in a study with Turkish language learners, Loewen et al. (2019) found participants performed better on the written sections of the test, namely writing, reading, and lexicogrammar sections, when compared to speaking and listening sections. More recently, Kessler et al. (2025) compared the efficacy of Babbel and Duolingo, and results showed improvement in various areas of language

https://doi.org/10.1017/S0272263126101521 Published online by Cambridge University Press

skills: Participants were more successful in listening, followed by speaking and word recognition sections, compared to the short-answer and error detection and correction sections. While these two studies show somewhat different patterns, the results may have been impacted by various ways to operationalize constructs and the different test formats.

Overall, previous research has suggested that using apps can result in language development in general, but the degree of efficacy for various language skills (e.g., receptive vs. productive) may differ. Accordingly, more fine-grained construct definitions and operationalizations are needed in the research domain of MALL. Furthermore, oral and written production-based skills have only been marginally explored in prior research, and no study has examined the development of pragmatic ability. Considering the importance of developing communicative competence during language learning, future research with MALL applications that target additional constructs is warranted. This expanded focus would provide a more holistic understanding of both the potentials and limitations of MALL applications particularly in developing oral proficiency and pragmatic competence, key areas that the present study aims to address.

Mobile-assisted language learning vs. classroom instruction

Previous research has compared MALL with classroom instruction to examine the comparability of learning outcomes between these two approaches. It is important to note, however, that these instructional contexts differ significantly, and methodologically, it is difficult to account for numerous intervening variables in a quasi-experimental design. As one of the earlier studies, Rachels and Rockinson-Szapkiw (2018) investigated L2 Spanish development among primary school students using a pretest/posttest design. Using a 50-item translation test, they found that Duolingo instruction was as effective as traditional classroom instruction in developing grammatical and vocabulary knowledge after studying 40 minutes per week for 12 weeks. Jiang et al. (2021) showed that Duolingo beginner-level learners of Spanish or French attained comparable L2 reading and listening skills when compared to university students studying the same L2s over four semesters. Although these studies provided insights into the role of mobile app-based language learning, researchers identified several critical limitations, including the lack of pretest/posttest comparison data using reliable measures, small sample sizes, and insufficient control over study time.

Recent research has sought to address some of these challenges. For instance, Rodríguez-Fuentes and Swatek (2023) compared university students in a non-credit-bearing 16-week language course with Duolingo users who were learning English in Colombia. They used the iTEP Academic Plus Test to measure development of grammar, listening, reading, speaking, and writing proficiency. Results showed Duolingo users improved in all skills except grammar, with gains surpassing those of students in the traditional classroom setting. Similarly, González-Fernández (2023) focused on overall proficiency and lexical development among English language learners in Spain. Both groups showed significant improvement in L2 proficiency and lexical knowledge after 16 weeks of study, demonstrating greater average gains in both receptive and productive vocabulary knowledge. However, when controlling for time of study, the Duolingo group had higher gains than the classroom instruction group in overall L2 proficiency, receptive grammar knowledge, and receptive vocabulary knowledge.

https://doi.org/10.1017/S0272263126101521 Published online by Cambridge University Press

Overall, empirical studies comparing app-based language learning and classroom instruction show that both contexts offer learning opportunities. In terms of their efficacy for specific language skills, the results are mixed. For instance, regarding vocabulary knowledge, while some research shows benefits of using Duolingo for receptive and productive vocabulary skills (e.g., González-Fernández, 2023), others found less robust evidence for productive vocabulary development (e.g., Jiang et al., 2021). Regarding listening skills, some research has reported limited benefits from app-based learning (González-Fernández, 2023; Loewen et al., 2019). However, the listening tests used in these studies did not focus on communicative competence. Although listening skills are critical for communicative competence in natural conversation contexts, minimal investigation has examined learning in contexts that resemble natural conversational interactions.

Engagement with mobile apps

When using MALL apps, learning can take place virtually anywhere, at any time (Jiang et al., 2020; Kukulska-Hulme & Shield, 2008; Li et al., 2023). In addition, users have different levels of engagement with the apps, such as the amount of time spent using the app (Jiang et al., 2020; Loewen et al., 2019), repetition of levels and corresponding lessons (García Botero et al., 2019; Isbell et al., 2017; Loewen et al., 2019), use of translations and grammar tips (Isbell et al., 2017; Vesselinov & Grego, 2012), and engagement with extra, non-mandatory activities (e.g., note-taking; Isbell et al., 2017; Loewen et al., 2019).

Among these measures, time spent on apps has been one of the most widely used constructs. For instance, Sudina and Plonsky (2024) examined the effects of frequency of logins, total duration, and intensity of using different content features of Duolingo (e.g., lessons, level reviews, skill practice, stories, and tests) over 6 months. The findings showed that the total minutes of app exposure positively correlated with written but not oral proficiency gains. Additionally, the number of lessons, level reviews, and login attempts was found to be strongly associated with proficiency gains. However, as Sudina and Plonsky noted, because the findings showed marginal gains, the significantly predictive variables should be interpreted with caution. Kessler et al. (2025) found a positive moderate correlation between participants' study time on apps and posttest scores after using Babbel or Duolingo for 8 weeks. Loewen et al. (2019) also reported that the amount of Babbel study time was the strongest predictor for learning Spanish, which was operationalized as vocabulary, grammar, and OPIc.

In a recent large-scale study, Hwang et al. (2024) examined MALL users' engagement profiles with Mango (a language learning application) and revealed three distinct patterns. Additionally, their findings suggested that the users with greater MALL acceptance showed higher degree of behavior engagement demonstrated through higher level of intense, frequent, and doable Mango app usage. They also demonstrated longer active usage and shorter breaks before logging into the app again, which suggests persistence of app use. Although Hwang et al. (2024) offer insightful information regarding MALL engagement, they do not offer discussion on the learning outcomes of such app use.

To summarize, previous literature suggests the potential of the use of MALL applications for language development. However, in order to understand the effectiveness of such language learning contexts more fully, research that offers information on the extent to which different language skills are learned through MALL applications is

https://doi.org/10.1017/S0272263126101521 Published online by Cambridge University Press

warranted. To understand which linguistic skills benefit the most, it is pertinent to examine various linguistic competencies including both receptive and productive skills within a single study and compare their effects. Additionally, since students may use MALL apps alongside traditional classroom learning, comparisons should include Duolingo-supplemented instruction. For broader generalizability of research findings, studies should also sample learners from multiple regions and classrooms.

Objectives of the present study

While reviewing previous research on mobile language learning applications, we identified several research gaps: (1) the need for a pretest/posttest design quasi-experimental study comparing Duolingo and classroom instruction, as well as a condition with a combination of the two; (2) the need to examine multiple linguistic skills in a single study using fine-grained and reliable measures that are appropriate for (true) beginners; (3) the importance of examining productive language skills in natural communicative contexts, and (4) the need for sampling practices that promote the generalizability of research. To address these gaps, this study aims to compare the efficacy of (1) Duolingo instruction (Duolingo-Only), (2) Classroom instruction (Classroom-Only), and (3) Classroom + Duolingo instruction on beginner-level French development. We operationalize French development across three language areas: productive vocabulary knowledge, grammatical competence, and pragmatic knowledge (*tu* vs. *vous*).

Additionally, we examined the amount of Duolingo usage and statistically controlled for French language proficiency at the pretest as well as L1 background (e.g., Romance language as L1 vs. Non-Romance L1s). The study was guided by the following research questions:

RQ 1: How do different learning contexts, Duolingo-Only, Classroom-Only, and Classroom + Duolingo, influence beginner-level L2 French language development?

RQ 2: How do amount of Duolingo use, L2 French proficiency, and L1 background moderate L2 French language development among beginner-level learners?

Previous mobile app efficacy research has characterized their studies as "natural experiments" which refer to "any event not under the control of a researcher that divides a population into exposed and unexposed groups" (Craig et al., 2017, p. 40). As described in Sudina and Plonsky (2024) and González-Fernández (2023), natural experiments recognize that researchers have less control over pedagogical interventions when compared to lab-based research. This lack of control is balanced by naturally occurring variation in exposure inherent to different intervention conditions (e.g., Duolingo use vs. classroom study). Because tight control of learner behavior is not possible in the conditions used in this study, we adopted a natural experiment approach with a full disclosure of uncontrolled factors as necessary limitations.

Method

Participants

Participants assigned to the Duolingo-Only group were invited to participate via email if they met the following criteria: (1) use of Duolingo via iOS, Android, or a web browser; (2) new user status (between 7 and 14 days of usage); (3) US-based IP address; (4) self-reported French beginner proficiency (true beginner); (5) no completion of units beyond

https://doi.org/10.1017/S0272263126101521 Published online by Cambridge University Press

Unit 4 in the Duolingo French course. A total of 134 participants who met the inclusion criteria and completed each step of the procedure (see Figure 3) were included in the initial sample. From these participants, we controlled for range of use to ensure comparability with Classroom and Classroom + Duolingo groups' time spent learning. Inspection of the initial distribution of total Duolingo usage (measured in hours) among the Duolingo-Only group indicated a right-skewed distribution with a relatively large standard deviation (M = 45.60, SD = 32.76), including several participants with very low and high total use: Min = 6.12, Max = 195.65. These data suggested the presence of outliers at both the top and bottom of the distribution. We used the interquartile range to identify the middle of the distribution and removed outliers outside this range. Specifically, we set a lower limit of 27 hours and an upper limit of 63 hours of Duolingo use during the study period. Of the 134 initial Duolingo participants who completed the study, 65 participants met these criteria and were included in the Duolingo-Only group for statistical analyses.

For the Classroom-Only and Classroom + Duolingo groups, the data were collected over two distinct semesters. Recruitment emails were sent to directors of French language programs and department chairs from 144 universities across the USA, inviting students to participate in the study based on their interest in or current use of Duolingo. Based on self-selection, 104 students were assigned to the Classroom-Only group and 203 students to the Classroom + Duolingo group. The latter group was also assigned to a virtual classroom in Duolingo Classrooms to track their usage. Of these participants, only those who completed all data collection procedures were included: 60 students in the Classroom + Duolingo group and 58 students in the Classroom-only group. The final sample represented 27 universities across 21 US states. Semester length ranged from 14 to 16 weeks, with credit hours ranging from 3 to 5.

Participant demographic information is provided in Table 1. Each sample showed a predominance of female participants. The Duolingo-Only group consisted primarily of university graduates who were, on average, older than participants in either classroom group.

Data instruments

The study employed a quasi-experimental design to compare French language development across three instructional conditions. Overall French language proficiency was measured, and language development was operationalized through five components: (1) productive vocabulary knowledge, (2) grammatical competence, (3) pragmatic competence (*tu* vs. *vous*), (4) communicative competence, and (5) oral complexity and fluency. Because the majority of our participants are (true) beginners, it was important to use instruments appropriate for their initial proficiency level and those that could capture any proficiency changes based on pretest and posttest score comparisons. Thus, we modified the following materials from previous research and/or designed new instruments with these goals in mind: (1) C-test, (2) picture description test, (3) error-correction test, and (4) multi-turn discourse completion test (MDCT). These instruments were developed through several rounds of pilot sessions. Reliability statistics for each measure are reported in the analysis section.

C-test: A measure of general L2 proficiency

C-tests consist of a short text (paragraph-length) in which the second half of every other word is deleted, requiring participants to add the missing word information (Raatz & Klein-Braley, 1981). C-tests are a validated general L2 proficiency measure in SLA

https://doi.org/10.1017/S0272263126101521 Published online by Cambridge University Press

Table 1. Participant characteristics

Characteristic	Duolingo-Only *n* (%)	Classroom-Only *n* (%)	Classroom + Duolingo *n* (%)
Age			
Mean	42.43	19.79	22.33
SD	13.19	2.51	7.49
Range	21–73	16–32	17–56
Gender			
Male	10 (15%)	12 (21%)	8 (13%)
Female	52 (80%)	44 (76%)	50 (83%)
Other	3 (5%)	2 (3%)	2 (3%)
First language(s)			
English	57 (88%)	43 (73%)	38 (64%)
English + Other(s)	8 (12%)	15 (26%)	22 (37%)
Romance Language as L1			
No	60 (92%)	52 (90%)	44 (73%)
Yes	5 (8%)	6 (10%)	16 (27%)
Other languages studied (except for French)			
No	28 (43%)	27 (47%)	27 (45%)
Yes	37 (57%)	31 (53%)	33 (55%)
Highest academic degree earned			
Some high school	0 (0%)	1 (2%)	0 (0%)
High school	8 (12%)	51 (88%)	39 (65%)
Associate degree	6 (9%)	3 (5%)	11 (18%)
2-year college	2 (3%)	0 (0%)	5 (8%)
4-year university	29 (45%)	2 (4%)	5 (8%)
Master's degree	15 (23%)	0 (0%)	0 (0%)
Ph.D.	1 (2%)	0 (0%)	0 (0%)
Other	4 (6%)	1 (2%)	0 (0%)
Classroom instruction format			
Face-to-face	n/a	50 (86%)	50 (83%)
Synchronous online	n/a	4 (7%)	9 (15%)
Asynchronous online	n/a	4 (7%)	1 (2%)

research (e.g., Counsell, 2018; Sudina & Plonsky, 2024). To ensure sufficient items for a reliable measure of general proficiency, we included three passages. The two passages we developed are provided in Figure 1 (21 gaps and 17 gaps, respectively). Since the second half of every other word was deleted, the specific target words were not pre-selected. However, when developing the passages, we paid close attention to the topics and linguistic features commonly found in introductory French instructional materials and curricula, namely, personal introduction, descriptions, preferences, and food. Minor adjustments were made to avoid unnecessary grammatical repetition (e.g., determiners and conjunctions). The final texts were piloted with learners and shared with experienced French instructors and Duolingo professionals to confirm their pedagogical appropriateness for this learner population.

Picture description test

To measure productive vocabulary knowledge gains, we created a picture description test, which has been widely used in SLA literature to elicit learner oral or written language output such as lexical complexity and syntactic complexity for various age groups and proficiency levels (e.g., Révész, 2009; Hwang, 2025). Pictures need to be

https://doi.org/10.1017/S0272263126101521 Published online by Cambridge University Press

Passage 1

Bonjour, je m'appelle Marise et j'ai un ami spécial ! I___ est bl___, brun e___, très pe___. Mon a___ a de___ belles orei_____ et el____ sont lon____. Le ma___, je l___ donne d___ la nourr___. Il ma___ des caro____ très rapid_____! Il n'a___ pas l___ chiens, ma___ il ad___ les enf___. Tu as deviné, mon ami c'est un lapin.

Passage 2

Coucou, tu es d'accord pour m'aider à organiser l'anniversaire de Clara ? La fê___ est sam___ chez m___. Nous av____ un exce___ gâteau a___ chocolat e___ mon fr___ a l___ décorations e___ la mus___. Viens av___ ton cop___. On v___ faire u___ sal___ de fru___ ensemble. Appelle-moi quand tu arrives.

Figure 1. C-test sample.

selected carefully to ensure familiarity with the content of pictures and cultural appropriateness, reducing cognitive load associated with content planning. We selected two comparable images of an outdoor park scene from BrainPOP ELL. As this instrument was used to measure productive vocabulary, we ensured high image quality including diverse characters varying in age, nationality, and gender performing various outdoor activities (e.g., running, riding a bicycle, playing music). Participants were given 60 seconds to study the picture and plan their response, and then had 3 minutes to describe the image. Because proficiency levels were low, participants were instructed to produce short sentences or to list words describing the characters, actions, and items in the pictures.

Error correction test

To measure grammatical competence, we created an error correction test, a common measure in SLA research (e.g., Smith et al., 2024). To inform item selection, we consulted introductory French textbooks, course syllabi from two US-based programs, DELF examination items, and Duolingo's A1 curriculum summary (personal communication, June 28, 2022). Originally a total of 30 items were created but after several rounds of pilot sessions and feedback from French instructors, the final test included 20 items. Error types included: subject-verb agreement ($n = 8$), gender agreement ($n = 3$), number ($n = 2$), partitive ($n = 2$), formality ($n = 2$); negation ($n = 1$), elision ($n = 1$), and inversion ($n = 1$). Error locations were controlled, with six errors at the start of the clause, seven each in middle and end positions. A higher number of items for subject-verb agreement was decided to account for different personal pronouns, regular/irregular verbs, and verb tense (present and passé composé), which is important part of beginner-level French grammar instruction. See sample items representing error types and location of errors:

**Ta frère travaille toute la journée au magasin.* (gender/Start)
*Le weekend, tu *regarde la télévision avec ta copine.* (subject-verb/Middle)
*Pour la salade, Julian prépare les tomates *rouge.* (number/End)

https://doi.org/10.1017/S0272263126101521 Published online by Cambridge University Press

Two test versions were created, varying only the content words while maintaining error types and locations in the sentence. Items were presented randomly, and test versions were counterbalanced between pretest and posttest to minimize test effects.

Multi-turn discourse completion test

We created two versions of a multi-turn discourse completion test (MDCT), each containing six scenarios to measure participants' appropriate use of *tu* vs. *vous* (our target pragmatic feature) and communicative competence (see Appendix B). Each scenario elicited five oral responses, yielding 30 unique responses per test version. The test required participants to produce 16 pragmatically appropriate structures (*tu* vs. *vous*) and 14 responses demonstrating their communicative competence. More specifically, we included two structures, the first (Type A) required the production of three unique questions to elicit pragmatically appropriate structures and two responses to demonstrate their communicative competence whereas the second (Type B) required the production of two unique questions and three responses. In total, 6 unique scenarios were created to elicit the target forms, which are listed in Table 2. Scenarios were presented randomly, and test versions were counterbalanced for the pretest and posttest to minimize test effects. This description was provided in English to the participants.

Each MDCT included nine images to simulate natural conversations: one image describing the scenario, five images prompting oral responses (white bubbles), and three images providing participant information (pink squiggly bubbles) (see Figure 2). For example, Figure 2 provides an MDCT of the Type A. Image 1 (top left) informed the participant of the character they would play, as well as the difference between oral production bubbles and information bubbles. Image 2 provided information written in English about the speakers (context and relationship) as well as the expected oral production. Image 3 prompted participants to listen to the oral input, provided in French. They could hear this a maximum of two times, since in a normal communicative context, asking for a repetition is acceptable. Image 4 asked the participant to make a question, Image 5 prompted the participant to listen to an answer (in French), Image 6 to make a question, Image 7 to listen to an answer and comprehend a question (in French), Image 8 an oral response to the question, and Image 9 an oral response, in the form of a comment or question. Type B differed slightly in that 2 questions were produced, and 3 responses were elicited. The test versions, counterbalanced between pretest and posttest, differed only in their images.

Table 2. Topic and description of the roles and situation

Topic	Description of the Roles and Situation
Learning French	Participant took the role of a classmate who had missed her French class. She called her friend to find out about the lesson/homework.
Baking a cake	Participant took the role of a young woman who calls her elderly neighbor, asking to borrow an ingredient.
Meeting a fellow traveler	Participant took the role of a young male traveler, meeting another young female traveler, and getting to know one another.
Meeting a new colleague	Participant took the role of a new employee, recently arriving at her new place of work. She is meeting a new colleague for the first time.
Getting to know a new roommate	Participant took the role of a young roommate who is asking about her new roommate's personalities, interests, and hobbies.
Learning French with a tutor	Participant took the role of a student, interacting with their new tutor who is an experienced language professor.

https://doi.org/10.1017/S0272263126101521 Published online by Cambridge University Press

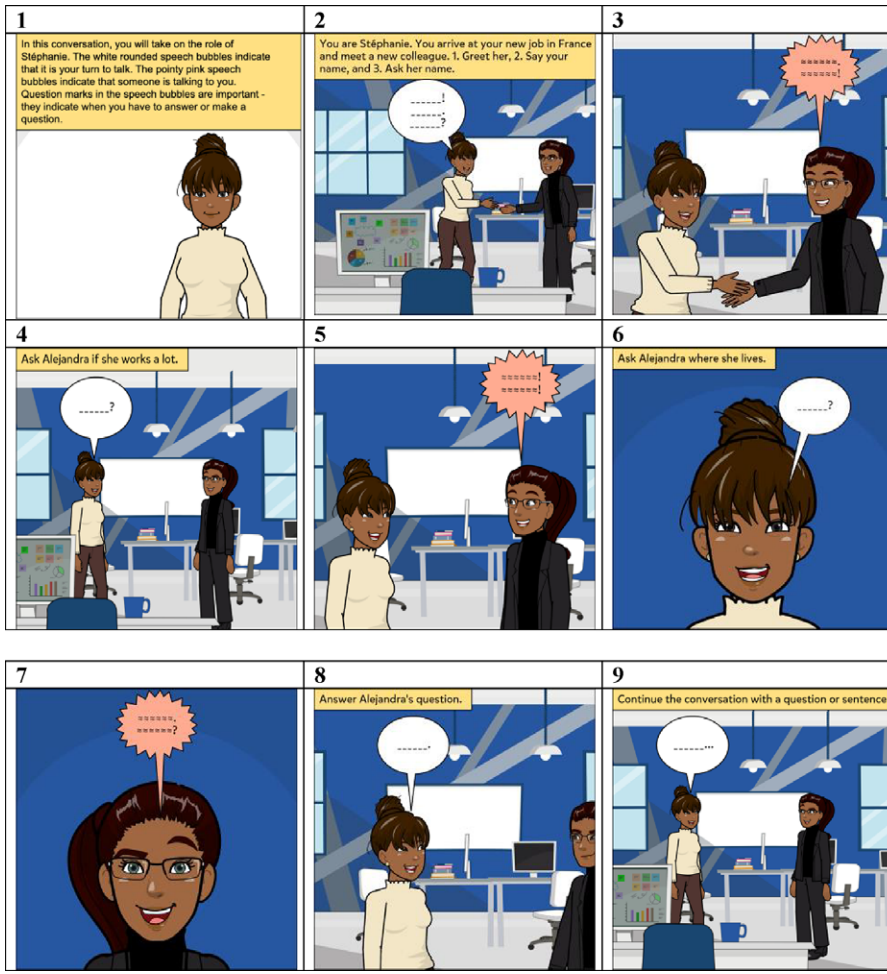


Figure 2. MDCT (Type A).
Note. The pink squiggly speech balloons indicate that participants will listen to audio responses.

User data

We obtained user data from the Duolingo for Schools platform to measure engagement levels for participants in the Duolingo-Only and Classroom + Duolingo groups. Engagement was operationalized as total time spent using Duolingo. From these data, we calculated average weekly Duolingo use.

Demographic survey

We collected demographic data from participants using an eight-part survey. In the current study, we analyzed data from the first section, which collected personal information, including nationality, first language, academic year, highest degree earned, age, and classroom instructional modality.

https://doi.org/10.1017/S0272263126101521 Published online by Cambridge University Press

Data collection procedure

To examine the influence of Duolingo on French language development, we recruited participants from two contexts: French language classrooms and Duolingo users in the USA. Recruitment strategies differed slightly between contexts (see Participants section for more details). Figure 3 summarizes the data collection procedure across all three groups.

Data coding

Participants completed four tests measuring different aspects of French language development, as described above. The following sections detail the analysis procedures for each test.

General French proficiency: C-test

For the C-test, each correct answer received 1 point, with total score used in the statistical analysis. Possible scores ranged from 0 to 55, corresponding to the total number of gaps across the three texts. For example, participants received 3 points for correctly completing the first sentence with *-te* (*fête*), *edi* (*samedi*), *oi* (*moi*).

La fê [ ] est sam [ ] chez m [ ].

Internal consistency of the C-test scores was high, with Cronbach's alpha values all approximately 0.98 when measuring C-tests individually at both pretest and posttest, and as a single combined test.

Productive vocabulary knowledge: Picture description test

To measure productive vocabulary knowledge gains, we calculated both the total number of French words produced during the picture description test (tokens) and

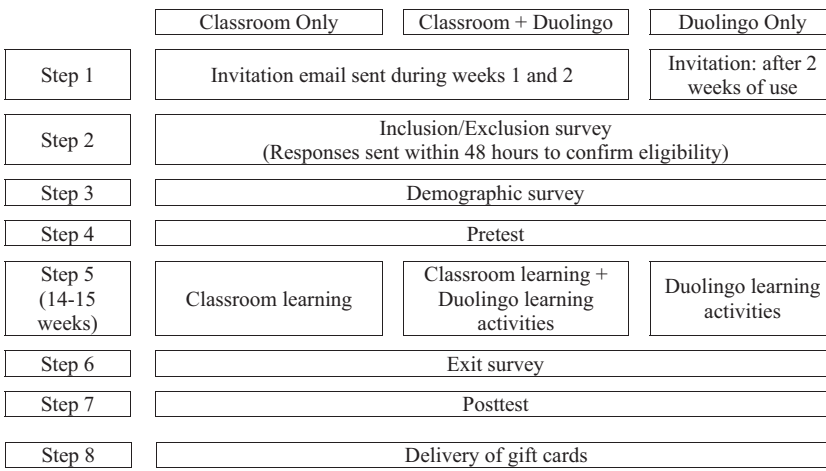


Figure 3. Data collection procedure.

https://doi.org/10.1017/S0272263126101521 Published online by Cambridge University Press

the number of unique words (types) in the picture description test. The change in word production from pretest to posttest provides an informative measure of growth for beginner-level learners, though it may be less suitable for advanced learners. Given the participants' beginning French proficiency level, an increase in word production and lexical variety was considered to reflect the development of foundational vocabulary knowledge. Two raters independently coded 20% of the data to establish interrater reliability. Agreement scores were calculated by dividing the total types of vocabulary from both raters and tokens of vocabulary. Agreement rates were high for both pretest (tokens: 97.8%, types: 96.8%) and posttest (tokens: 97.6%, types: 97.0%).

Grammatical competence: Error correction test

For the error correction test, participants received 1 point for correctly identifying each error and 1 point for accurately correcting the error, yielding a maximum of 2 points per item. For example, in the item, *Il joue avec une balle dans le maison*, the error involves the article gender. Participants earned 1 point for identifying the error by writing either *le* or *le maison*. For error correction, participants earned 1 point for writing *la*, *la maison*, *une*, or *une maison*. One of the 20 items, *je appeler*, allowed partial points for both error identification and error correction. Participants could receive partial points for identifying the missing reflexive particle (*je me*/*se appeler*) and/or correctly forming one element (e.g., *je m'appeler* or *je appelle*). Test scores ranged from 0 to 40 points. Internal consistency reliability was high, with Cronbach's alpha values of 0.95 or higher when measuring different test versions at both pretest and posttest, and as combined tests.

Pragmatic competence and communicative competence: MDCT

For each MDCT, participants produced contextually appropriate questions or responded to oral prompts. Pragmatic competence was measured through appropriate use of the second-person marker, *tu* vs. *vous* (informal/formal). Each pragmatically appropriate form received 1 point, for a maximum of 16 points. Communicative competence was operationalized as participants' ability to use the target language appropriately in conversational interactions (Whyte, 2019). The MDCT responses were scored based on their logical continuation of the conversation. Each appropriate response received 1 point, yielding a maximum of 14 points on the communicative competence measure. Two independent coders rated 20% of the pretest and posttest data. For the pretest, the agreement levels were 95.01% for pragmatic competence and 95.98% for communicative competence. For the posttest, the agreement was 95.23% for pragmatic competence and 98.56% for communicative competence.

Statistical analysis

Pretest-posttest gains among groups

To assess group improvements from pretest to posttest, we fitted linear mixed-effects regression models using R version 4.4.0 (R Core Team, 2024) and the *lme4* package version 1.1.35.3 (Bates et al., 2015). Each model used French language ability as the dependent variable, with a three-way interaction among Group (Classroom-Only, Duolingo-Only, and Classroom + Duolingo), Test (Pretest and Posttest), and total pretest C-test scores. This interaction enabled assessment of both within-group gains and between-group differences in gains, while controlling for initial proficiency levels.

https://doi.org/10.1017/S0272263126101521 Published online by Cambridge University Press

The models also included a categorical variable for Romance language L1 status to account for potential language transfer effects.

The models included random intercepts for subjects to account for by-participant variation, thereby distinguishing individual performance variation from predictor variable effects. We did not fit a random intercept for items because we used aggregated scores at pre and post, and the random intercepts for items consistently led to singular fits (i.e., no meaningful variance to measure). The R syntax for these models was as follows:

$$lmer(\text{dependent variable} \sim \text{Group}_{(\text{Classroom Only,Duolingo-Only,Classroom+Duolingo})} * \text{Time}_{(\text{Pretest,Posttest})} * \text{C-Test}_{(\text{numerical,z-scored})} + \text{RomanceL1}_{(\text{yes,no})} + (1|\text{Subject})$$

Using the *emmeans* package (version 1.10.1; Lenth, 2024), pairwise comparisons were conducted between pretest and posttest scores within groups to assess significant growth. Between-group comparison of gains was then conducted to determine differential improvement patterns. The analysis also examined the effects of pretest proficiency and Romance L1 status.

Effects of Duolingo use

To examine the effects of Duolingo usage on posttest performance (RQ 2), additional models were fitted using data from Duolingo-Only and Classroom + Duolingo participants. The models included a three-way interaction among Group, Time, and weekly Duolingo minutes. Using the *emtrends* function from the *emmeans* package, we examined whether Duolingo usage significantly influenced posttest scores and pretest-to-posttest gains. The R syntax was as follows:

$$lmer(\text{dependent variable} \sim \text{Group}_{(\text{Duolingo-Only,Classroom+Duolingo})} * \text{Time}_{(\text{Pretest,Posttest})} * \text{Minutes Per Week}_{(\text{numerical,z-scored})} + \text{C-Test}_{(\text{numerical,z-scored})} * \text{Time}_{(\text{Preteset,Posttest})} + \text{RomanceL1}_{(\text{yes,no})} + (1|\text{Subject})$$

Continuous predictor variables were z-scored (standardized to a mean of zero and standard deviation of one), enabling cross-scale comparisons and standardized coefficient interpretation in terms of standard deviations from the mean. Model fit was assessed using the *performance* package (version 0.11.0; Lüdecke et al., 2021), which provided conditional and marginal R^2 values to measure variance explained by random and fixed effects combined, and fixed effects alone. Due to the large number of comparisons across multiple models, an alpha value of 0.005 was selected to control for Type I errors.

Audio quality issues and technical problems prevented accurate scoring of some responses. Consequently, sample sizes across models varied from the total participant pool. Each pretest and posttest analysis included only participants with valid scores at both time points (pretest and posttest). The results section reports the actual sample size (n) for each group along with descriptive statistics for each outcome measure.

Results

Before answering the research questions, analyses examined two variables that potentially influenced gains across test measures: Duolingo usage patterns and C-test performance.

https://doi.org/10.1017/S0272263126101521 Published online by Cambridge University Press

Duolingo usage data

Table 3 presents the means and standard deviations for weekly Duolingo usage. The Duolingo-Only group averaged 3.5 times more minutes per week than the Classroom + Duolingo group. The Classroom + Duolingo group showed variance nearly equal to its mean, while the Duolingo-Only group's variance was approximately 25% of its mean. This difference partially reflects the trimming of extreme values in the Duolingo-Only group and indicates more variable usage patterns among Classroom + Duolingo participants.

General French proficiency: C-test data

Table 4 presents means and standard deviations for C-test scores, which measured general proficiency. Group averages were above zero on the pretest in all three groups. These scores indicated pre-existing French proficiency at treatment onset, supporting its inclusion as a covariate in the regression models.

Statistical comparisons of pretest and posttest scores indicated all three groups demonstrated significant gains. Predicted gains were similar across groups, showing increases of 9–10 points from pretest to posttest (see Table 5). The magnitude of gains did not differ significantly across groups. No significant differences emerged among groups at either testing point. These results suggest comparable initial French proficiency across groups, as measured by the C-test. All groups showed similar improvements in French proficiency at posttest (see Appendix C). Romance L1 status showed

Table 3. Average number of minutes spent using Duolingo per week

Group	*M* (*SD*)	Median	Range
Duolingo-Only	117.72 (31.00)	112.68	80.94–216.08
Classroom + Duolingo	45.74 (39.46)	33.26	3.23–221.89

Table 4. Descriptive statistics for pretest and posttest C-test scores

	Pretest			Posttest		
Group	*M* (*SD*)	Median	Range	*M* (*SD*)	Median	Range
Classroom-Only	11.26 (9.34)	9	0–33	20.36 (11.00)	19	0–40
Duolingo-Only	13.53 (8.28)	11	0–39	22.75 (7.67)	22.5	0–45
Classroom + Duolingo	12.29 (8.62)	10.5	0–32	22.34 (11.11)	24.5	0–46

Note: C-test scores ranged from 0 to 55. Sample sizes (*n*) were lower than total participant numbers due to incomplete pretest/posttest data. Classroom-Only *n* = 58 (of 58); Duolingo-Only *n* = 64 (of 65); Classroom + Duolingo *n* = 56 (of 60).

Table 5. Pretest-posttest gains in C-test scores

Within-group differences from pretest to posttest					
Group	Estimate	95% CI	SE	*t*	*p*
Classroom-Only	9.10	[7.15, 11.05]	0.99	9.21	<.001
Duolingo-Only	9.22	[7.36, 11.08]	0.94	9.80	<.001
Classroom + Duolingo	10.05	[8.07, 12.04]	1.01	10.00	<.001

https://doi.org/10.1017/S0272263126101521 Published online by Cambridge University Press

no significant effects: estimate = 2.59, 95% CI [–1.20, 6.37], SE = 1.92, t = 1.348, p = 0.179. Models examining weekly Duolingo usage effects (measured in minutes per week) on posttest scores showed no significant influence for either the Duolingo-Only or Classroom + Duolingo group (see Appendix C). The conditional R^2 was 0.756 and the marginal R^2 was 0.216, indicating fixed effects explained 20% of the variance, with substantial variance explained by the random effects structure (random intercept for participants).

Productive vocabulary knowledge (types and tokens): Picture description task data

To measure gains in productive vocabulary knowledge, oral production was analyzed for vocabulary types and tokens during a picture description task. Table 6 presents means and standard deviations for French vocabulary tokens and types. All groups showed increased production of both vocabulary tokens and types from pretest to posttest.

Two multilevel regression models for tokens and types revealed statistically significant differences (all p <.001). All groups showed significant increases in tokens and types at posttest, producing approximately 25 to 32 more tokens and 13 to 16 more types. Total word production and vocabulary range increased significantly from pretest to posttest across all groups. Table 7 presents within-group pretest to posttest contrasts. No significant differences emerged between groups in the magnitude of change, suggesting similar increases in types and tokens across all groups (see Appendix C).

The two models also revealed that increased C-test scores correlated with significantly higher posttest scores for tokens and types across all groups. C-test scores showed no significant effect on pretest-to-posttest gains for any group (see Appendix C). Romance L1 status showed no significant effects for tokens (p = .757) or types (p = .732). In terms of model fit, the conditional and marginal R^2 for the tokens model were .677 and .403, and for the types model were .667 and .498. Fixed effects accounted for approximately 40% and 50% of the variance in the tokens and types models, respectively, with additional variance explained by random effects (random intercept for participants).

Table 6. Produced tokens and types in French during the picture description task

Tokens				
	Pretest		Posttest	
Group	*M* (*SD*)	Range	*M* (*SD*)	Range
Classroom-Only	21.35 (25.65)	0–143	49.18 (36.31)	0–167
Duolingo-Only	30.11 (24.04)	0–110	55.33 (28.21)	11–127
Classroom + Duolingo	23.92 (26.32)	0–145	55.52 (39.11)	11–199
Types				
	Pretest		Posttest	
Group	*M* (*SD*)	Range	*M* (*SD*)	Range
Classroom-Only	12.41 (10.84)	0–42	26.61 (14.06)	0–61
Duolingo-Only	17.89 (11.61)	0–57	31.08 (13.10)	8–62
Classroom + Duolingo	13.89 (11.25)	0–50	30.10 (15.66)	9–89

Note: Sample sizes (n) were lower than total participant numbers due to incomplete pretest/posttest data: Classroom-Only n = 49 (of 58); Duolingo-Only n = 63 (of 65); Classroom + Duolingo n = 52 (of 60).

https://doi.org/10.1017/S0272263126101521 Published online by Cambridge University Press

Table 7. Pretest-posttest gains in number of produced French tokens and types (productive vocabulary knowledge)

Tokens					
Group	Estimate	95% CI	SE	*t*	*p*
Classroom-Only	28.30	[20.65, 35.95]	3.88	7.31	<.001
Duolingo-Only	25.28	[18.52, 32.05]	3.43	7.38	<.001
Classroom + Duolingo	31.82	[24.40, 39.23]	3.75	8.48	<.001
Types					
Group	Estimate	95% CI	SE	*t*	*p*
Classroom-Only	14.38	[10.93, 17.83]	1.75	8.23	<.001
Duolingo-Only	13.16	[10.11, 16.21]	1.55	8.52	<.001
Classroom + Duolingo	16.25	[12.91, 19.59]	1.69	9.60	<.001

Additional models examining weekly Duolingo usage effects (measured in minutes per week) showed no significant influence on posttest productive vocabulary scores for either group. Duolingo usage showed no significant effects on pretest-to-posttest gains for tokens. However, increased Duolingo use correlated with stronger gains in types, but only for the Duolingo-Only group (p = .003). Each standardized increase in average Duolingo minutes corresponded to approximately 7.6 more types on the posttest beyond the average pretest-to-posttest gains. Full contrasts are provided in Appendix C.

Grammatical competence: Error correction test data

To measure gains in grammatical competence, participants completed an error correction test. Table 8 presents means and standard deviations for error spotting and correction. All groups showed improved ability to spot and correct grammatical errors in French sentences from pretest to posttest.

Table 8. Descriptive statistics for error spotting and correction test scores

Spotting Errors				
	Pretest		Posttest	
Group	*M (SD)*	Range	*M (SD)*	Range
Classroom-Only	5.12 (3.46)	0–14.5	9.94 (4.82)	1–19
Duolingo-Only	5.85 (3.87)	0–17	10.72 (3.78)	1–17
Classroom + Duolingo	4.36 (3.13)	0–13.5	9.80 (5.44)	0–20
Correcting Errors				
	Pretest		Posttest	
Group	*M (SD)*	Range	*M (SD)*	Range
Classroom-Only	2.92 (3.29)	0–13	8.02 (4.85)	0–18
Duolingo-Only	4.18 (3.78)	0–17	8.78 (3.84)	2–17
Classroom + Duolingo	2.38 (2.97)	0–12.5	8.72 (5.51)	0–19

Note: Sample sizes (n) were lower than total participant numbers due to incomplete pretest/posttest data. Classroom-Only n = 57 (of 58); Duolingo-Only n = 65 (of 65); Classroom + Duolingo n = 58 (of 60). Maximum possible scores: Error spotting = 20, Error correction = 20.

https://doi.org/10.1017/S0272263126101521 Published online by Cambridge University Press

Table 9. Within-group contrasts in growth from pretest to posttest for error spotting and correction

Spotting Errors					
Group	Estimate	95% CI	SE	*t*	*p*
Classroom-Only	4.86	[3.75, 5.97]	0.56	8.64	<.001
Duolingo-Only	4.97	[3.93, 6.01]	0.53	9.41	<.001
Classroom + Duolingo	5.46	[4.36, 6.55]	0.56	9.84	<.001
Correcting Errors					
Group	Estimate	95% CI	SE	*t*	*p*
Classroom-Only	5.13	[4.10, 6.16]	0.52	9.83	<.001
Duolingo-Only	4.69	[3.73, 5.66]	0.49	9.58	<.001
Classroom + Duolingo	6.38	[5.36, 7.39]	0.51	12.40	<.001

Two multilevel regression models for error spotting and correction revealed statistically significant differences (all p <.001). All groups showed significant increases at posttest, spotting 4.8 to 5.5 more errors and correcting 4.7 to 6.4 more errors. Table 9 presents within-group pretest to posttest contrasts. No significant differences emerged between groups in the magnitude of change, indicating similar improvement in error spotting and correction errors across groups (see Appendix C).

Higher C-test scores were associated with significantly higher posttest scores for error spotting and correction across all three groups (all p <.001). Conversely, there was no significant effect of C-test scores on gains between pretest and posttest for any group. Furthermore, no significant differences were observed between participants who spoke a Romance L1 for either error spotting (p = .330) or correction (p = .069). In terms of model fit, the conditional and marginal R^2 for error spotting model were 0.628 and 0.522, respectively. For the error correction model, the conditional and marginal R^2 were 0.689 and 0.545, respectively. This indicates that the fixed effects structure for both models accounted for 52% and 55% of the variance, with the random effects structure (random intercept for participants) capturing additional unexplained variance.

Additional models testing the effects of Duolingo usage (measured in minutes per week) on the error correction tests found no significant effects of Duolingo use on posttest scores or on gains from pretest to posttest for the Duolingo-Only and Classroom + Duolingo groups. This suggests that variation in Duolingo use did not have a strong effect on error spotting and correction on the Error Correction Task (see Appendix C).

Pragmatics (tu vs. vous) and communicative competence: Multi-turn discourse completion test (MDCT)

Participants completed the MDCT to assess their pragmatic knowledge, particularly regarding the use of the informal (*tu*) and formal (*vous*) forms in various communication contexts, as well as their overall communicative competence. The means and standard deviations for communicative and pragmatic competence (*tu* vs. *vous*) on the MDCT are presented in Table 10. These data indicate that, on average, all groups scored higher on the MDCT measures for pragmatic and communicative competence in the posttest compared to the pretest.

Two multilevel regression models were used to predict pragmatic competence (*tu* vs. *vous*) and communicative competence. These differences were statistically

https://doi.org/10.1017/S0272263126101521 Published online by Cambridge University Press

Table 10. Descriptive statistics for pragmatics competence and communicative competence scores on MDCT

Pragmatic Competence (tu vs. vous)				
	Pretest		Posttest	
Group	*M (SD)*	Range	*M (SD)*	Range
Classroom-Only	7.11 (3.34)	0–14	8.30 (2.93)	0–15
Duolingo-Only	7.66 (1.94)	2–11	9.10 (1.86)	4–14
Classroom + Duolingo	6.00 (3.56)	0–13	9.47 (3.34)	0–16
Communicative Competence				
	Pretest		Posttest	
Group	*M (SD)*	Range	*M (SD)*	Range
Classroom-Only	5.47 (3.26)	0–12	7.98 (2.68)	0–12
Duolingo-Only	6.25 (2.31)	1–11	8.87 (2.15)	3–12
Classroom + Duolingo	5.37 (2.97)	0–12	8.61 (2.83)	0–12

Note: Sample sizes (*n*) were lower than total participant numbers due to incomplete pretest/posttest data. Classroom-Only $n = 57$ (of 58); Duolingo-Only $n = 61$ (of 65); Classroom + Duolingo $n = 57$ (of 60). Maximum possible scores: Pragmatic competence = 16; Communicative Competence = 14.

significant (all $p < .005$). On average, all groups showed significant increases in both pragmatic competence (ranging from scores 1.1 to 3.4 points) and communicative competence (ranging from 2.4 to 3.2 points) from pretest to posttest. This suggests that participants in all groups improved their scores on both competence measures from pretest to posttest, as shown in Table 11.

A comparison of gains revealed no significant difference in communicative competence scores among the groups. However, significant differences were found for pragmatic competence (*tu* vs. *vous*) scores. Specifically, the Classroom + Duolingo group had significantly higher gains from pretest to posttest than both the Classroom-Only and Duolingo-Only groups (both $p < .001$). These differences were comparable in magnitude, with gains being approximately two points higher when compared to the Classroom-Only and Duolingo-Only groups. No other significant between-group contrasts were observed; full contrasts are provided in Appendix C.

Table 11. Within-group contrasts in growth from pretest to posttest for pragmatic and communicative competence during MDCT

Pragmatic Competence (tu vs. vous)					
Group	Estimate	95% CI	SE	*t*	*p*
Classroom-Only	1.13	[0.41, 1.85]	0.37	3.09	.002
Duolingo-Only	1.48	[0.78, 2.19]	0.36	4.17	<.001
Classroom + Duolingo	3.46	[2.74, 4.18]	0.36	9.50	<.001
Communicative Competence					
Group	Estimate	95% CI	SE	*t*	*p*
Classroom-Only	2.44	[1.84, 3.04]	0.30	8.05	<.001
Duolingo-Only	2.68	[2.10, 3.26]	0.30	9.10	<.001
Classroom + Duolingo	3.22	[2.62, 3.81]	0.30	10.64	<.001

https://doi.org/10.1017/S0272263126101521 Published online by Cambridge University Press

Higher C-test scores were associated with higher pragmatic competence (*tu* vs. *vous*) scores for the Classroom-Only ($p < .001$) and Classroom + Duolingo group ($p = .001$), but not for the Duolingo-Only group ($p = .249$). A similar pattern was observed for communicative competence scores in the posttest (all $p < .001$). In contrast, no significant effect of C-test scores on gains from pretest to posttest was found for any group (see Appendix C). Additionally, no significant differences in pragmatic competence (*tu* vs. *vous*) scores were observed based on Romance L1 status ($p = 0.35$). However, a significant effect was found for communicative competence scores, with participants who spoke a Romance L1 receiving ratings approximately 1.43 points higher than those who did not speak a Romance L1 (95% CI [0.57, 2.29], SE = 0.44, $t = 3.28$, $p = .001$). This effect was consistent across pretest and posttest. Regarding model fit, the conditional and marginal R^2 values were 0.62 and 0.31 for the pragmatic competence model and 0.73 and 0.45 for the communicative competence model. This indicates that the fixed effects structure accounted for 31% and 45% of the variance, respectively, with additional variance captured by the random effects structure (random intercept for participants).

Further models testing the impact of Duolingo usage (measured in minutes per week) on MDCT scores found no significant effects on posttest scores or the gains from pretest to posttest for the Duolingo-Only and Classroom + Duolingo groups. This suggests that variation in Duolingo usage did not substantially influence communicative or pragmatic competence scores (see Appendix C).

Summary of the findings

Table 12 summarizes the main findings from the current study, organized by each target construct examined.

Discussion

The purpose of the study was to measure and compare the development of French language ability among beginning learners over the span of a university semester in the United States across three instructional conditions: (1) Classroom-Only; (2) Duolingo-Only; and (3) Classroom + Duolingo. Using a quasi-experimental pretest-posttest design, we investigated five aspects of language competence: (1) overall French proficiency, (2) productive vocabulary ability, (3) grammatical competence, (4) pragmatic competence (*tu* vs. *vous*), and (5) communicative competence. The first research question focuses on the three groups, and the second research question addresses three moderating variables: Duolingo use (the total amount of usage per week), French proficiency (C-test results at the pretest), and L1 background (Romance language vs. other languages). Drawing from existing theory and prior SLA studies, we created bespoke assessment materials specifically designed to measure these dimensions of language ability for (true) beginners.

Our results demonstrate positive effects for all three instructional contexts. All three groups exhibited significant improvement in their French language ability as measured through multiple dimensions of L2 competence. This finding was expected, particularly because we are dealing with learners of limited to no pre-existing French proficiency. As long as an instructional modality provides adult learners with access to target language input, language learning is expected to occur (Van Patten et al., 2020). Based on our results, all three modalities were sufficient to accomplish this goal. The finding that Duolingo can support language development for various linguistic dimensions including both receptive and productive skills is important given the limited number of

https://doi.org/10.1017/S0272263126101521 Published online by Cambridge University Press

https://doi.org/10.1017/S0272263126101521 Published online by Cambridge University Press

Table 12. Summary of the main findings

Construct	Task	Operationalization	Pretest-Posttest Gains	Group Differences	Duolingo Use	French Proficiency	Romance L1
Overall proficiency	C-test	C-test scores	Yes, all groups	No	No	N/A	No
Productive vocabulary knowledge	Park description task	Total tokens produced	Yes, all groups	No	No	Higher posttest scores	No
		Total types produced	Yes, all groups	No	Duolingo-Only Group: more usage = higher gains	Higher posttest scores	No
Grammatical competence	Error correction task	Scores for spotting the errors	Yes, all groups	No	No	Higher posttest scores	No
		Scores for correcting the errors	Yes, all groups	No	No	Higher posttest scores	No
Pragmatic competence (tu/vous)	MDCT	Scores on MDCT items for *tu* vs. *vous*	Yes, all groups	Classroom + Duolingo gains > Classroom-Only and Duolingo-Only gains	No	Higher posttest scores for Classroom-Only & Classroom + Duolingo	No
Communicative competence	MDCT	Scores on MDCT items for adequate responses	Yes, all groups	No	No	Higher posttest scores	Higher scores (across tests)

studies that have examined the effectiveness of such technology using a pretest-posttest research design (Loewen et al., 2019).

A consistent pattern emerges when comparing the degree of growth among the groups. Our results revealed no significant differences between the Classroom-Only and Duolingo-Only groups for any of our measures. The extent of the gains from pretest to posttest between these two groups was comparable, with neither group performing significantly better than the other. One might expect to see different performance between Duolingo users and classroom learners on measures depending on curricula, particularly whether the instruction is focused more strongly on receptive or productive skills. Yet we found comparable performance between the groups for both productive and receptive knowledge, including productive vocabulary, grammatical knowledge, the targeted pragmatic feature (*tu* vs. *vous*), and communicative competence. Such findings expand the current body of literature attesting to the benefits of mobile app-based language learning (e.g., Smith et al., 2024; Sudina & Plonsky, 2024). The Classroom + Duolingo group demonstrated comparable learning outcomes across all measures, except for one measure of pragmatic competence focusing on learning when to use *tu* vs. *vous*, where they showed significantly higher gains than the other two groups.

Two important observations from the dataset warrant consideration when comparing learning gains among these groups. First, the amount of time-on-task and instruction may differ among the groups. Because we adopted a natural experimental approach in this study (see Sudina & Plonsky, 2024), the findings should be interpreted accordingly, reflecting both practical differences in learning conditions and higher ecological validity. Additionally, although the amount of time is one way to measure the amount of engagement with the instructional materials (i.e., behavioral engagement), it does not necessarily indicate the depth of processing of the input and cognitive engagement with the instructional materials (Hwang et al., 2024). Thus additional measures for the engagement with instructional materials need to be considered.

Another caution that needs to be highlighted here is the heterogeneity of learner backgrounds when comparing the participants from the Duolingo-Only and Classroom-Only groups. Similar to previous mobile app-based research (e.g., Jiang et al., 2021), mobile app users who managed to fulfill the study inclusion requirements in terms of app usage amount are often older than typical university students, and, more importantly, the majority have already obtained a university degree or higher. Therefore, the participants in the Duolingo-Only group of our final dataset may not be representative of the broader Duolingo population (e.g., individuals from diverse age groups or who sporadically use the application). However, it is also important to note the current attrition trend reported in previous MALL research. For instance, Sudina et al. (2025) identified age as a significant predictor of attrition in the use of Duolingo. Similarly, Hwang et al. (2024) also reported that older Mango app users primarily used a single language learning app persistently compared to the younger users, indicating a higher attrition rate. Such difference in learner backgrounds must be taken into account when interpreting our findings, especially since age factor might be associated with other learner characteristics such as maturity and language learning motivation (Sudina et al., 2025). Nevertheless, this research provides valuable data and offers insights regarding age effects in L2 acquisition (Birdsong, 2018). Older language learners in the Duolingo-Only condition demonstrated improvement in language skills using mobile apps, and the degree of their learning effects is comparable to that of university students enrolled in the first semester of French language instruction.

In our study, there was no significant difference among the three groups except for the pragmatic competence measures focusing on *tu* vs. *vous*, where the Classroom +

https://doi.org/10.1017/S0272263126101521 Published online by Cambridge University Press

Duolingo Group made stronger gains than the other two groups. One might expect that learners using a combined approach may achieve better learning outcomes compared to those using only Duolingo or only taking French classes due to additional instructional time. However, the findings showed that the relative use of Duolingo for participants in the Classroom + Duolingo group was rather limited. This might reflect the reality that since the course curriculum and Duolingo curriculum are not aligned, students may not view Duolingo as a supplement to enhance their classroom-based French learning. Instead, they may have seen Duolingo primarily as an additional task competing for time in their already busy schedules. Alternatively, if students viewed Duolingo mainly as an opportunity to practice, but found the content did not align with their French curriculum, they may have failed to see its benefits and felt that the practice was not useful for passing their course (leading to lower use). Given these factors, it becomes less surprising that the amount of Duolingo use was not found to be a significant moderating factor of French language learning for the Classroom + Duolingo group. Considering the current trend in the use of MALL apps for various purposes includes entertainment, knowledge building, and learning new skills (such as language learning; e.g., Enayat et al., 2025), future research should explore how language learning mobile applications can effectively supplement classroom instruction, particularly when classroom curricula integrate apps like Duolingo to reinforce classroom instructional content. In terms of the Classroom + Duolingo condition's outperformance on the pragmatic competence measure, it is important to note that this group also started with much lower pretest scores when compared to the other two groups. Despite these different starting points (Classroom $M = 7.11$; Duolingo Only $M = 7.66$, Classroom + Duolingo $M = 6.00$), all three groups showed significant improvement and converged to similar post-test levels (ranging from 8.30 to 9.47). These statistically larger gains are thus attributable to the Classroom + Duolingo group's lower starting point. The possibility remains that the combined instruction helped to overcome the lower starting point, but this is mostly speculative and would require more testing of balanced and imbalanced starting points to fully determine.

We also included initial French proficiency (operationalized as pretest C-test scores) in our models, and the findings suggested that higher proficiency at the beginning of the experiment was associated with higher posttest scores. However, in terms of gain scores, having a higher degree of French knowledge did not facilitate higher gain scores. Considering our participants were (true) beginner learners of French, this may suggest that French proficiency level variance at the pretest might not be significant enough to show any statistical impact on gain scores. Another account could be that at the beginner level, French proficiency level differences are not playing a major role in these three learning contexts.

In the current study, because of the linguistic similarities among Romance languages such as Spanish and French, we controlled participants' L1 in our statistical modeling, which made the models more robust. The findings suggested that, in general, Romance L1 was not a moderating factor in the study except for the communicative competence. Although Romance Language backgrounds such as Spanish did not seem to benefit learning French grammar, productive vocabulary knowledge, or the use of *tu* vs. *vous* through Duolingo or classroom instruction, it did help the participants for learning how to carry out oral conversation. This might be because learning grammar, vocabulary, and pragmatic features at this level might be more influenced by the amount of input and learner cognitive processing of those features in the instruction. However, this result needs to be interpreted with some caution because of unbalanced number of participants with and without Romance Language backgrounds across the three groups.

https://doi.org/10.1017/S0272263126101521 Published online by Cambridge University Press

In terms of the role of the amount of Duolingo use, there was a lack of usage effects for the Duolingo-Only group. More Duolingo use was only associated with stronger productive vocabulary knowledge for this group. However, it should be noted that we only included participants who used Duolingo for a total of 27 to 65 hours over the study—usage effects might be stronger with a wide range of Duolingo usage in the final dataset. The relative lack of effects for usage above and beyond these gains suggests the amount of time spent on Duolingo (behavior engagement) might not be directly associated with the degree of learning due to other moderating factors. Also some language features like productive vocabulary might be more sensitive to the exposure time (behavior engagement) than others. Notably, as mentioned above, while we measured engagement with Duolingo through the average number of minutes of Duolingo use per week, this metric cannot tell us about the depth of processing while using the application, an important aspect of usage data. Additionally, since users can repeat lessons without proceeding to new ones, high usage time may not correspond to progression through the curriculum. Overall, regarding the effect of Duolingo usage, our findings differed from previous research suggesting that the amount of MALL application use is associated with learning (Kessler et al., 2025; Loewen et al., 2020). To gain a clearer picture of application engagement, future research should consider the multidimensional aspects of MALL engagement (Hwang et al., 2024), including frequency (frequency of logins), density (duration per login), and content progression (number of units completed) as presented in Sudina and Plonsky (2024).

Our study included some methodological improvements over prior studies that better capture the actual extent of change in language skills. First, by using and counterbalancing comparable pretest and posttest measures, we assessed language gains across different linguistic competencies including both receptive skills and oral language production, an underexplored construct in this research. The findings further attest to the appropriateness of our assessment materials for beginning learners. Our challenge was to create comparable pretest and posttest measures, knowing that most of our learners would have little to no existing knowledge of French. While our results suggest that some participants either had prior knowledge or were adept at intuiting the rules of French, our use of pretest scores allowed us to account for these baseline differences. This approach provided a more fine-grained measure of growth from pretest to posttest, rather than only measuring groups at posttest.

Second, in this study, we controlled for prior knowledge and L1 background. We used a C-test to measure general French language proficiency, which is a common language competence measure in SLA research (Eckes et al., 2006). The findings indicate that all three groups demonstrated significant gains between the pretest and posttest on the C-test, and there were no significant differences among the three instructional conditions. Although we screened participants for prior knowledge by asking them to indicate their current proficiency in the target language, the pretest C-test results showed some variations that we needed to attend to. Subsequently, we used the pretest C-test scores as a measure of learner proficiency to control for prior knowledge in all remaining statistical models for various linguistic learning outcomes.

Third, beyond overall proficiency and grammatical competence, we used oral production tests to measure language gains, which have been less widely implemented in previous MALL application research. For instance, despite the high reliability and validity of standardized tests, capturing beginning-level learners' language gains over one semester using such measures can be challenging. The measures used in this study were generally motivated by prior SLA research and successfully identified the degree of gains across the three groups for various dimensions of L2 competence. Specifically, we

https://doi.org/10.1017/S0272263126101521 Published online by Cambridge University Press

examined several under-explored constructs in MALL research, including pragmatic competence (*tu* vs. *vous*) and communicative competence using oral production data. Because mobile application studies often focus on beginning-level learners, as they are the most popular learner group, future research should continue to explore appropriate measures for this population.

Despite the contributions of this research, several limitations need to be acknowledged. First and foremost, although we put our best effort into fair sampling for generalizability of our research, a number of unavoidable sampling-related concerns have arisen. For instance, a considerable amount of attrition raises a concern that participants included in the final data set are less representative of the population. It is also worth commenting on the noticeable difference in age among our three groups, in that both university groups were on average younger than the Duolingo-Only group (M = 19, 22 vs. M = 42). Such a large age difference may cause other confounding factors such as maturity and motivation. Age was not a primary question of ours; however, the final pool of the data after data screening in our research suggests age may be a relevant consideration for future research in this area. MALL research might benefit from more systematic examination of age effects or stratified sampling approaches to better isolate the variables of primary interest, especially as age may correlate with different groups of language learners engaging with apps such as Duolingo.

Secondly, comparing the three instructional conditions in this report involves several critical confounding variables, which are related to the nature of the natural experiment discussed in the literature review section. Although our inclusion of students from a wide range of university French programs enhances the generalizability of the results, the differences in curricula among programs, as well as the programs' characteristics (i.e., an exact amount of instruction time per semester, target language coverage in course materials, teaching methods), were not fully taken into consideration. Additionally, regarding the Classroom + Duolingo group, our initial plan was to have course instructors require Duolingo use to accurately operationalize our intended instructional condition (using Duolingo as supplemental material as a part of university-level course curricula). However, due to the wide geographic distribution of schools and French programs, such an arrangement was not feasible.

Furthermore, although studies have operationalized engagement with language apps in various ways such as frequency, duration, and intensity (Loewen et al., 2019, 2020; Sudina & Plonsky, 2024), this study used the average amount of time spent on Duolingo per week. Other studies using the same measure have reported moderate correlations between the amount of time spent on Duolingo and learning gains (e.g., Loewen et al., 2019), suggesting that other ways to model app use should be considered in future research. Finally, participants completed the tests online at their chosen locations for practicality. Although we requested that participants complete the tests in quiet locations with breaks between test sections, we could not ensure compliance, meaning that some learners may have spent more or less time on the tests and completed them in different environments. Additionally, because participants used their own laptops and tablets, there was some data loss due to poor-quality audio files or missing recordings. Such limitations were unavoidable as the study was conducted remotely across the country.

Conclusion

Despite the presence of intervening variables inherent in a natural experiment, the adoption of a pretest/posttest quasi-experimental research design allows this study to offer insights into the extent to which each instructional condition promotes language

https://doi.org/10.1017/S0272263126101521 Published online by Cambridge University Press

gains in various areas. It is encouraging to see how all three groups showed improvement over the course of one semester on various linguistic competencies. It is particularly noteworthy that Duolingo instruction is equivalent to classroom instruction for developing beginner-level learners' oral production, particularly regarding facets of communicative competence and the appropriate use of *tu* vs. *vous*. This finding is especially significant given that the Duolingo users' backgrounds ranged widely in terms of the purpose of language learning and language learning backgrounds.

Mobile use is a part of essential daily routines and its technologies have emerged as an important mediating tool for L2 teaching and learning. Because mobile users could access language learning applications easily with flexibility in terms of time and location, it may be an efficient tool with affordable cost. The findings of the study suggest its effectiveness in learning various language skills, which was comparable to classroom instruction. While this study did not examine learner characteristics among Duolingo users, future research should consider these learner characteristics to fully understand who benefits from using Duolingo and how it could be used to maximize learning. We chose university-level first-semester French classes to operationalize classroom instruction, and thus our participants in classroom conditions shared a homogeneous background (e.g., age, taking the class partially to meet their graduation requirements). Therefore, it is not appropriate to conclude that Duolingo instruction is comparable to *all* types of classroom instruction for *all* learners, especially since we could not control the classroom instruction in this study. However, since the goal of Duolingo is to offer free and inclusive education to all, our data suggests this goal is being met for adult learners of all backgrounds, not just at the university level. To make fair comparisons, future MALL application efficacy research should explore similar groups of students when comparing the efficacy of a language learning application (e.g., Duolingo) to classroom instruction. Furthermore, as classroom instruction varies across settings—including different pedagogical approaches (e.g., content-based instruction, task-based instruction), instructional modalities (online vs. face-to-face vs. hybrid), and different age groups (e.g., middle school)—classroom instruction should be defined more narrowly and clearly.

The ultimate goal of language learning is to use the target language in diverse social contexts and engaging in meaningful interactions with others. Thus, MALL application efficacy research should prioritize measuring communication skills in real-life contexts. Thanks to the recent surge of using AI in designing MALL applications, incorporating production-based and interaction-driven language practice is becoming more feasible than ever and technologies within language apps change rapidly.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S0272263126101521.

Data availability statement. The data will be available upon request.

Acknowledgements. The project was funded by the Duolingo Efficacy Research Program. We are grateful for Dr. Xiangying Jiang for her support throughout the project. Special thanks go to the editor, Dr. Luke Plonsky and anonymous reviewers who have shared insightful comments. We are also indebted to French program directors and instructors who helped us to recruit participants as well as our participants.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

https://doi.org/10.1017/S0272263126101521 Published online by Cambridge University Press

Birdsong, D. (2018). Plasticity, variability and age in second language acquisition and bilingualism. *Frontiers in Psychology*, *9*, 81. https://doi.org/10.3389/fpsyg.2018.00081

BrainPOP ELL. https://educators.brainpop.com/brainpop-esl-teacher-resources/brainpop-esl-picture-prompts/

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

Counsell, C. L. (2018). The C-test in French: Development and validation of a language proficiency test for research purposes. In J. M. Norris (Ed.), *Developing C-tests for estimating proficiency in foreign language research* (pp. 205–232). Peter Lang.

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*(1), 39–56. https://doi.org/10.1146/annurev-publhealth-031816-044327

Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing*, *23*, 290–325. https://doi.org/10.1191/0265532206lt33

Enayat, M. J., Ghadim, N. A., & Arabmofrad, A. (2025). Effects of two mobile-assisted language learning apps on L2 receptive and productive vocabulary knowledge: A mixed-methods study. *System*, *133*, 103763. https://doi.org/10.1016/j.system.2025.103763

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. (2023). *The effectiveness of Duolingo vs. classroom instruction on Spanish speakers' L2 English proficiency and lexical development*. Duolingo Research Report.

Hwang, H. (2025). Growth of lexical and syntactic complexity, accuracy, and fluency in spoken production of first language and second language children. *System*, *132*, 103695.

Hwang, H, 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*, *46*(4), 969–995. https://doi.org/10.1017/S0272263124000354

Isbell, D. R., Rawal, H., Oh, R., & Loewen, S. (2017). Narrative perspectives on self-directed foreign language learning in a computer- and mobile-assisted language learning context. *Languages*, *2*(4), 1–20. https://doi.org/10.3390/languages2020004

Jiang, X., Chen, H., Portnoff, L., Gustafson, E., Rollinson, J., Plonsky, L., & Pajak, B. (2021). *Finishing half of B1 on Duolingo comparable to five university semesters in reading and listening*. Duolingo Research Report.

Jiang, X., Rollinson, J., Plonsky, L., & Pajak, B. (2020). *Finishing A2 on Duolingo comparable to four university semesters in reading and listening*. Duolingo Research Report.

Kessler, M., Loewen, S., & Gönülalc, T. (2025). Mobile-assisted language learning with Babel and Duolingo: Comparing L2 learning gains and user experience. *Computer Assisted Language Learning*, *38*(4), 690–714. https://doi.org/10.1080/09588221.2023.2215294

Kittredge, A., Hopman, E., Reuveni, B., Dionne, D., Freeman, C., & Jiang, X. (2025a). Mobile language app learners' self-efficacy increases after using generative AI. *Frontiers in Education*, *10*, 1499497. https://doi.org/10.3389/feduc.2025.1499497

Kittredge, A., Lee, C., & Jiang, X. (2025b). *Video Call improves Japanese English learners' speaking skills*. Duolingo Research Report.

Kittredge, A., Peters, R., Neumann, F., & Jiang, X. (2024). *Duolingo learners can start a conversation after 4-6 weeks of app use*. Duolingo Research Report.

Kukulska-Hulme, A., & Shield, L. (2008). An overview of mobile assisted language learning: From content delivery to supported collaboration and interaction. *ReCALL*, *20*(3), 271–289. https://doi.org/10.1017/S0958344008000335

Lenth, R. (2024). emmeans: Estimated Marginal Means, aka Least-Squares Means (1.10.1) [Computer software]. https://CRAN.R-project.org/package=emmeans

Li, R., Zou, D., Reynolds, B.L., & Vazquez-Calvo, B. (2023). Editorial: Mobile assisted language learning: Developments, affordances, and solutions. *Frontiers in Psychology*, *14*, 1–2. doi:10.3389/fpsyg.2023.1293483

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

https://doi.org/10.1017/S0272263126101521 Published online by Cambridge University Press

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

Lüdecke, D., Ben-Shachar, M., 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

R Core Team. (2024). R: A language and environment for statistical computing (4.4.0) [Computer software]. https://www.R-project.org/

Raatz, U. & Klein-Braley, C. (1981). The C-Test – a modification of the cloze procedure. In T. Culhane, C. Klein-Braley, D. K. Stevenson (Eds.), *Practice and problems in language testing*, University of Essex.

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

Renaud, C. (2010). *On the nature of agreement in English-French acquisition: A processing investigation in the verbal and nominal domains (Unpublished doctoral dissertation)*. Indiana University, Bloomington.

Révész, A. (2009). Task complexity, focus on form, and second language development. *Studies in Second Language Acquisition*, *31*(3), 437–470. https://doi.org/10.1017/S0272263109090366

Rodríguez-Fuentes, R. A., & Swatek, A. (2023). *A comparison between classroom and MALL instruction with Duolingo: Learning English at the A2 CEFR level*. Duolingo Research Report.

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

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

Sudina, E., Teimouri, Y., & Plonsky, L. (2025). L2 grit and age as predictors of attrition in mobile-assisted language learning. *Learning and Individual Differences*, *120*, 102704. https://doi.org/10.1016/j.lindif.2025.102704

VanPatten, B., Keating, G. D., & Wulff, S. (Eds.). (2020). *Theories in second language acquisition: An introduction* (3rd ed.). Routledge.

Vesselinov, R., & Grego, J. (2012). *Duolingo effectiveness study: Final report*. Queens College, City University of New York.

Vesselinov, R., & Grego, J. (2016). *The Babbel efficacy study: Final report*. Queens College, City University of New York.

Whyte, S. (2019). Revisiting communicative competence in the teaching and assessment of language for specific purposes. *Language Education & Assessment*, *2*(1), 1–19. https://doi.org/10.29140/lea.v2n1.33

Cite this article: Kim, Y., Payant, C., Skalicky, S., & Namkung, Y. (2026). Comparing the effectiveness of Duolingo, Classroom instruction, and Classroom + Duolingo instruction conditions on beginner-level French language development. *Studies in Second Language Acquisition,* 48: 680–707. https://doi.org/10.1017/S0272263126101521

https://doi.org/10.1017/S0272263126101521 Published online by Cambridge University Press