Computer Assisted Language Instruction Consortium

CALICO JOURNAL (ONLINE) ISSN 2056–9017

Article

The Effectiveness of Duolingo English Courses in Developing Reading and Listening Proficiency

Xiangying Jiang[1], *Ryan Peters*[2], *Luke Plonsky*[3], *and Bozena Pajak*[4]

Abstract

This study evaluated the reading and listening proficiency levels of 245 English language learners who completed the Basic content (through Common European Framework of Reference for Languages [CEFR] level A2) of one of the following three Duolingo English courses: English for Japanese, Spanish, or Portuguese speakers. Participants self-reported having little to no prior proficiency in English and used Duolingo as their only learning tool. Their language skills were measured using the reading and listening sections of the STAMP 4S English Test from Avant Assessment. The results show that, on average, learners at the end of the Basic content (A2) scored Intermediate High in both reading and listening in all three courses. Given that Duolingo's English courses are CEFR-aligned with an expected proficiency outcome of Intermediate Mid upon completion of the A2 content, it is noteworthy that participants across the three studied courses surpassed these expectations by achieving an additional American Council on the Teaching of Foreign Language (ACTFL) sublevel in both reading and listening. Such results present a robust and coherent body of

Affiliations

[1] Duolingo, Pittsburgh, Pennsylvania, USA.
email: xiangying@duolingo.com
[2] Duolingo, Pittsburgh, Pennsylvania, USA.
email: ryanpeters@duolingo.com
[3] North Arizona University, USA.
email: luke.plonsky@nau.edu
[4] Duolingo, Pittsburgh, Pennsylvania, USA.
email: bozena@duolingo.com

Received: 24 July 2023 Accepted after revision: 21 January 2024

CALICO JOURNAL https://doi.org/10.1558/cj.26704

evidence affirming the efficacy of Duolingo's English courses in enhancing learners' competencies in reading and listening comprehension.

Keywords: Duolingo; efficacy; English; reading proficiency; listening proficiency; MALL.

1. Introduction

Mobile-assisted language learning (MALL) has become increasingly popular in recent years, thanks in part to the ubiquity of smartphones and tablets. With the ability to access language learning apps and resources at any time and from anywhere, mobile devices have transformed the way people approach language acquisition. This flexibility allows for more consistent and frequent exposure to the target language, which is crucial for successful language acquisition. Moreover, many mobile language learning apps are designed with gamification elements that make the learning process more enjoyable and motivating, leading to increased user engagement.

The unprecedented growth of online and mobile-based language courses offered by both educational institutions and commercial organizations has caught the attention of scholars and practitioners as well as the general public. However, the effectiveness of these courses is still met with skepticism by numerous language teaching professionals (Brown, 2023; Lin & Warschauer, 2015). This skepticism is mainly attributed to the scarcity of empirical evidence that demonstrates the proficiency of learners through standardized and validated language proficiency measures (Burston & Giannakou, 2022; Tarone, 2015). One of the goals of the present study is, therefore, to describe the ability of app-based language learners in terms of standardized tests and their corresponding proficiency levels.

Research on the effectiveness of such language learning tools also contributes to ongoing discussions about the roles of factors such as input, instruction, and context in second language acquisition (SLA). Online and mobile-based language learning represents a unique context that transcends borders and can accompany learners in virtually any setting (Loewen et al., 2020), challenging the traditional categorization of learning context as classroom versus laboratory and second versus foreign language learners. In the scenario of mobile-based language learning, instruction goes beyond classrooms to a virtual setting where learners interact with course content and gamified features autonomously. The fact that learners can study at their own convenience for a few minutes a day without being confined to a fixed class schedule can also provide evidence on the theoretical arguments regarding the effects of spaced

versus massed instruction (Carpenter, 2020; Rohrer, 2015; Sudina & Plonsky, 2023; see also meta-analytic evidence on the effects of distributed practice in Kim & Webb, 2022).

Among the studies that have researched the effectiveness of MALL, English has been the dominant target language (Burston & Giannakou, 2022). In recent decades, the global landscape of English language learning has witnessed a significant surge in interest and engagement. As the lingua franca of international communication, business, and academia, English has solidified its position as an indispensable tool for individuals seeking to navigate the increasingly interconnected world. This growing demand for English proficiency has spurred the development of various innovative pedagogical approaches, ranging from online learning platforms to immersive language programs. Despite these advances, disparities in access to quality education and resources persist, underscoring the need for a more equitable distribution of opportunities for learners across the globe. As a freely and universally available resource, Duolingo's mission on educational equity and accessibility seeks to meet this need.

Understanding the effectiveness of popular language learning apps is of particular importance in the field of MALL research due to such apps' pervasiveness and accessibility. Take Duolingo as an example. As a language learning platform that offers free online courses (with optional paid subscriptions) on the web and mobile apps, Duolingo holds approximately 90% of global active users in the commercial app-based language learning market (M Science, 2023, p. 5). By making its content free and universally available, it reduces the disparities in access to quality education. As a result, the effectiveness of language learning apps such as Duolingo has piqued the interest of teachers, administrators, parents, learners, and researchers alike. Such stakeholders may reasonably question, however, the efficacy of Duolingo in terms of language gains. Therefore, the present study aims to evaluate the effectiveness of Duolingo's English courses. Specifically, we set out to assess the listening and reading proficiency levels of Duolingo learners in three English courses using standardized assessments. These learners, who self-reported using Duolingo as their sole English-learning tool, were tested after completing the Basic content (through Common European Framework of Reference for Languages [CEFR] A2) in Duolingo's English course for Japanese, Spanish, or Portuguese speakers.

2. Literature Review

2.1 Research on the Effectiveness of MALL

Since the appearance of MALL-oriented research in the early 1990s, thousands of primary studies have been carried out and published, in addition

to many overviews and meta-analyses. In a recent meta-analysis, which employed a strict set of inclusion criteria, Burston and Giannakou (2022) found substantial effects for MALL interventions for both between-group designs (i.e., samples of mobile-based learners compared to non-mobile learners; $k = 84$) and within-group designs (i.e., pre–post designs; $k = 56$): $g = 0.72$ and $g = 1.16$, respectively. The magnitude of these effects is comparable to those observed in other domains of instructed SLA (Plonsky, 2017).

Despite evidence of the effectiveness of MALL, several design-related weaknesses have been noted which call into question the findings in this domain. Burston and Giannakou (2022), for example, criticized MALL researchers for routinely relying on very small samples. Concerns over low statistical power and the threat it poses for internal and, hence, external validity, have been expressed on multiple occasions in instructed SLA (e.g., Loewen & Hui, 2021; Norouzian, 2020; Plonsky, 2013). Being aware of these concerns, we have intentionally collected a larger sample in this study to overcome the problems associated with small sample sizes.

A second major concern is instrumentation. More specifically, as noted by Loewen et al. (2020), most studies of MALL effectiveness appear to employ researcher-made tests to assess the effectiveness of the instruction learners receive. Such tests are convenient and very common throughout the field (Park et al., 2022; Thomas, 2006), but they often lack the validity evidence needed to draw meaningful conclusions. As Norris and Ortega (2012) put it, "problematic … is the tendency to assume—rather than build an empirical case for—the validity for whatever assessment method is adopted [in L2 research]" (pp. 574–575). The use of non-standardized tests also greatly limits our ability to understand mobile learners' development in relation to widely understood proficiency scales, such as those of the American Council on the Teaching of Foreign Language (ACTFL) and the CEFR. However, there is a small but growing body of studies that have tested the effectiveness of MALL using standardized assessments, as also found in the present study.

2.2 Research on the Effectiveness of Duolingo

Shortt et al. (2023) conducted a systematic review of Duolingo-related studies published between 2012 and 2020. Looking across their sample ($K = 35$), the authors observed an overall positive relationship between the use of Duolingo and language performance. More specifically, Duolingo has been found to be effective in developing learners' ability in multiple linguistic domains in English, as well as other languages, including vocabulary (Ajisoko, 2020), listening skills (Bustillo et al., 2017), and communicative skills (Rolando et al., 2019). More nuanced findings and insights on the

effects of Duolingo can be gleaned by examining individual studies that cover various target skills, which we will now discuss in detail.

Loewen et al. (2019), for example, examined gains in multiple skills made in a sample of nine Duolingo learners of Turkish as a foreign language. Their scores on the posttest, which was based on an "in-house" exam for a beginner-level university Turkish course, varied across skill areas, but tended to be stronger in written (reading, writing) than oral skills (listening, speaking). After an average of 29 hours learning Turkish on Duolingo during one semester, only one participant received what would be considered a passing grade based on a test designed for a university-based Turkish course. However, Shortt et al. (2023) cautioned on the accuracy of these findings due to the small sample size, potentially confounding variables, and a lack of standardized assessments.

In contrast to Loewen et al.'s (2019) interest in more global proficiency gains, Ajisoko (2022) investigated the effect of Duolingo in improving one particular skill in English—reading comprehension. A total of 15 engineering undergraduates were asked to earn 20 Experience Points (XPs) per day on Duolingo on weekdays for a period of two months. Although there was no control group in the design, the pretest and posttest reading comprehension scores showed large gains (d = 2.02). Fatmawati et al. (2023) were, likewise, interested in a single target domain—grammar. A group of participants learning with Duolingo was compared to another group learning with a different application, Cake, in acquiring the grammatical feature of the simple present tense. The study followed a comparison group design with pre- and posttests, and all participants (N = 58) were middle school students in Indonesia. The results of the study showed that learners who used Duolingo had a significantly greater gain in their knowledge of the simple present tense than those who used Cake.

Several other studies have compared the effectiveness of Duolingo with face-to-face instruction or other learning tools. Overall, such studies have found the effects of Duolingo to be comparable or, in some cases, superior to other tools and types of instruction. Rachels and Rockinson-Szapkiw (2018), for example, compared the use of Duolingo with traditional classroom instruction. The authors employed a pretest–posttest design to compare face-to-face Spanish classroom instruction with Duolingo's Spanish course for English speakers in an elementary school. Students from six third- and fourth-grade classes used Duolingo to learn Spanish, while the other five classes attended regular face-to-face Spanish classes (N = 164). Both groups learned Spanish for 40 minutes per week for 12 weeks, after which they were assessed on Spanish vocabulary and grammar using multiple-choice items. The same test was used in both the pretest and posttest. The researchers found no significant difference between the two groups and concluded that Duolingo was as good as face-to-face classes for teaching Spanish to elementary students.

Similarly, James and Mayer (2019) found no significant differences between college students ($N = 64$) who learned Italian on Duolingo during seven sessions and those who learned it using an online slideshow, which indicated that Duolingo was similarly effective in comparison with other commonly used tools. Kessler et al. (2023) compared the effectiveness of Duolingo with Babbel, another commercial language learning app. A group of 59 adult learners studied Turkish using Babbel ($n = 27$) or Duolingo ($n = 32$) for eight weeks and were then tested on their Turkish language skills: reading, listening, writing, speaking, vocabulary, and grammar. The results indicated that both groups developed their language skills, but there were no statistically significant differences. Again, these findings confirmed that Duolingo was as effective as other tools.

The number of research publications on Duolingo's effectiveness has been growing over the years. In addition to those studies conducted by independent researchers such as those mentioned so far, Duolingo's Efficacy Research Lab has also been conducting studies to investigate the app's efficacy. For instance, the study by Jiang et al. (2021) was conducted by Duolingo internal researchers with an external collaborator. The study reported the reading and listening proficiency outcomes of Duolingo learners after completing the Basic content of its Spanish and French courses, up to level A2 based on the CEFR (Council of Europe, 2001). The study focused on Duolingo learners who self-reported having little to no prior knowledge in the target language and used Duolingo as their only learning tool. Results from the ACTFL Reading Proficiency Test (RPT) and Listening Proficiency Test (LPT) demonstrated that on completion of the A2 content, these students reached Intermediate Low in reading comprehension and Novice High in listening comprehension. These results were comparable to university students who completed four semesters of Spanish or French courses in university language programs (Winke et al., 2020).

Looking across this body of research, there seems to be a growing collection of work to show the effects of MALL in general and, in particular, Duolingo, both on its own and in comparison with other platforms and learning contexts. At the same time, many studies on the efficacy of Duolingo, including those reviewed above, possess one or more of the aforementioned weaknesses; namely, a lack of standardized instruments and small sample sizes. The current study addresses these problems by using a standardized assessment and a much larger sample size.

2.3 The Present Study

The goal of this study was to investigate the reading and listening abilities of Duolingo learners in three English courses—English for Japanese speakers

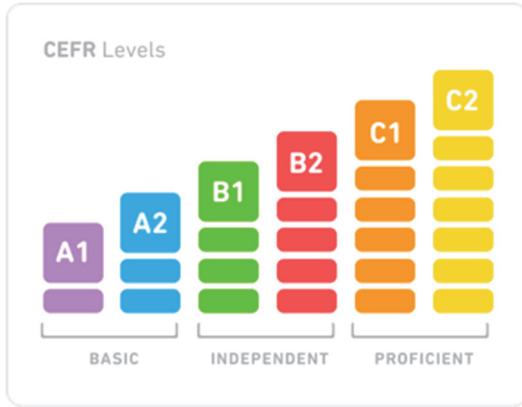

Figure 1: A diagram of the six CEFR levels.

(EN<JA), English for Spanish speakers (EN<ES), and English for Portuguese speakers (EN<PT)—after learners had finished the first four sections of their respective courses. The Duolingo English curriculum is aligned with the CEFR, which is an international language proficiency standard that defines learning goals for Basic (A1–A2), Independent (B1–B2), and Proficient (C1–C2) users (see Figure 1; Council of Europe, 2001). The first four Duolingo course sections cover the material through to level A2.

The Duolingo course structure is organized based on sections and units. There are four sections in the Basic content: a brief intro section, two sections of A1 content, and a longer section of A2 content. Each section has a different number of units. An average unit contains around 25 2–3-minute sessions. Table 1 shows the number of units in each section of the CEFR Basic content. To reach the end of the A2 content, learners need to complete a total of 80 units in the course for Japanese speakers, and 81 units in the courses for Spanish

Table 1:
CEFR Basic Content in the Duolingo English Courses

CEFR Basic level	Duolingo course section	Number of units in EN<JA	Number of units in EN<ES and EN<PT
Pre-A1	Section 1	5 units	10 units
A1.1	Section 2	17 units	19 units
A1.2	Section 3	18 units	16 units
A2	Section 4	40 units	36 units
Total		80 units	81 units

and Portuguese speakers. While these three English courses are all aligned to the same standard, the exact content is different because each Duolingo course is constructed with the specific language background (interface language) in mind. In our previous analysis with another course, it takes about 122 hours to reach the end of the A2 content if learners start from the very beginning of the course (56 hours to complete the A1 content and 66 hours to complete the A2 content). These Duolingo English courses also include content beyond A2, but in this study learners were assessed at the end of A2.

The content in each unit includes lessons that either introduce new material or review previously covered content, as well as short stories. Lessons include several activity types targeting vocabulary, grammar, reading, listening, writing, and speaking. To facilitate listening and speaking development, Duolingo provides learners with many opportunities to listen to the target language and speak it out loud. All English course content is accompanied by audio, and learners are allowed to play the audio at varied speeds as often as they need. In addition, speech recognition technology is used for all speaking exercises, in order to provide learners with feedback. Review lessons provide personalized spaced repetition of the material to ensure each learner practices their weaker areas. Finally, short stories provide discourse-level reading and listening comprehension practice, reinforcing and enriching learners' knowledge by situating the lesson content in everyday contexts.

The current study investigated the reading and listening proficiency outcomes of learners in three Duolingo English courses (English for Japanese, Spanish, and Portuguese speakers) after the participants completed the first four sections of their course (through CEFR A2).

The research question was as follows: *what levels of reading and listening proficiency do participants achieve when they reach the end of the Basic content (A2) in Duolingo's English course for Japanese, Spanish, or Portuguese speakers?*

3. Methods

3.1 Participants

The study participants comprised a total of 245 English learners—81 Japanese speakers, 72 Spanish speakers, and 92 Portuguese speakers—in their corresponding Duolingo English courses. To qualify for study participation, learners had to meet four criteria related to age, language background, course content completion, and use of non-Duolingo learning materials.

3.1.1 Age

All the study participants were 18 years of age or older. It was explicitly stipulated in the initial invitation that they must be at least 18 years of age to partake in the study. Participants' ages were also confirmed by self-reported age in the background survey.

3.1.2 Language Background

All participants had to have limited prior proficiency in English. English was not their home language before age 6 (as self-reported in the background survey), and their English proficiency at the time of starting to learn it on Duolingo was low (as self-reported on the app, which is information that Duolingo collects from all learners when they start a new course for the purposes of learner analytics).

For learners in the EN<ES and EN<PT courses, self-reported prior proficiency in English had to be 0–2 on a 0–10 scale, in order to participate in the study, where 0 represents "I have no knowledge of English at all," and 10 indicates "I have perfect knowledge of English." This proficiency scale was used before Duolingo transitioned from its original tree structure homescreen to the current path structure homescreen, at which point the scale changed. Table 2 shows the percentage of study participants at each self-reported proficiency level prior to instruction.

The learners in the EN<JA course started learning on Duolingo after its homescreen transition, so they responded to a new four-level prior proficiency scale instead. Their self-reported prior proficiency had to be either "I don't know any English" or "I know basic words and phrases," in order to participate in the study. The two higher levels of the scale were "I can understand simple conversations" and "I am intermediate or advanced." Except for two participants who reported knowing no English, all the participants from the EN<JA course reported knowing basic words and phrases.

Table 2:
Percentage of Participants at Each Proficiency Level Prior to Instruction

Prior proficiency scale (0–10)	**EN<ES (*N* = 72)**	**EN<PT (*N* = 92)**
0	18%	10%
1	32%	30%
2	48%	60%

Table 3:
Participants' Course Starting Point after Placement

CEFR A1 level	Duolingo course section	EN<JA (*N* = 81)	EN<ES (*N* = 72)	EN<PT (*N* = 92)
Pre-A1	Section 1	37%	85%	82%
A1.1	Section 2	30%	11%	12%
A1.2	Section 3	33%	4%	6%

3.1.3 Duolingo Course Completion

All participants completed the A2 material in their corresponding Duolingo English courses. This meant that their latest completed session in that course had to be within units 78–80 in the course for Japanese speakers and units 79–81 in the courses for Spanish and Portuguese speakers. Please note that these participants were naturalistic Duolingo learners who happened to be at the end of the A2 content during the data collection period. Because most participants started the courses with some knowledge of English, they began to learn at different places in the courses based on their performance in the placement test.[1] The placement test data showed that all participants started learning on Duolingo in the A1 sections. Table 3 shows the percentage of learners who started at each A1 section.

3.1.4 Use of Non-Duolingo Learning Materials

All participants self-reported not taking other English classes or using other apps or programs to study English while they were learning English on Duolingo.

3.2 Instruments

3.2.1 The Background Survey

The background questionnaire included questions related to participants' language background, reasons for learning the language, highest level of education, age group, and whether they took classes or used other apps/programs to learn English during the time they learned English on Duolingo. The answers to some of these questions confirmed eligibility for participation (see section 3.1 Participants). We note, however, that the survey did not ask about informal exposure to English through movies, friends, or social media, for example. The background survey was translated into Japanese, Spanish, and Portuguese for accessibility purposes. The responses to the survey questions are summarized in the Appendix.

3.2.2 The STAMP 4S English Test: Reading and Listening Sections

The reading and listening sections of the STAMP 4S English Test were used to assess the participants' proficiency in these skill areas for this study. The STAMP 4S English Test is a commercial standardized test provided by Avant Assessment. The acronym STAMP stands for "Standards-Based Measurement of Proficiency," and 4S refers to the four skills of reading, writing, listening, and speaking. The STAMP 4S English Test is an online, ACTFL-aligned, computer-adaptive test of English language proficiency accredited by the American Council on Education (ACE).

The reading and listening sections of the test consist of 30 multiple-choice questions each, which assess test takers' ability to comprehend a variety of written or spoken texts used for general communicative purposes in English. Each reading and listening question has an associated benchmark level. Test takers complete questions at various levels because the reading and listening sections are computer-adaptive. Novice items target sub-skills of topic identification and recognition of high-frequency words, phrases, and simple sentences. Intermediate items target sub-skills of comprehension of main ideas and supporting details on more familiar topics, and understanding facts, details, and specific information. Advanced items target sub-skills of comprehension of main ideas and supporting details on less familiar and more complex topics, understanding facts, details, and specific information, inferring indirect information from the context, and paraphrasing.

The two sections of the test take 60–75 minutes to complete. The test is automatically scored and test results are reported in two ways: through ordinal ratings on a scale of 1–9 (STAMP levels) and through interval scaled scores. According to Avant Assessment (Santos, 2022), the internal consistency reliability coefficients (Cronbach's alpha) of the reading and listening sections of the STAMP 4S English Test are 0.89 and 0.90, respectively.

Both the STAMP level ratings and scaled scores are aligned with three broad levels on the ACTFL proficiency scale: Novice, Intermediate, and Advanced, with each being further divided into Low, Mid, and High (ACTFL, 2012). As shown in Figure 2, the STAMP scale of 1–9 is aligned to nine ACTFL sublevels:

STAMP Level 1	STAMP Level 2	STAMP Level 3	STAMP Level 4	STAMP Level 5	STAMP Level 6	STAMP Level 7	STAMP Level 8	STAMP Level 9
Novice			Intermediate			Advanced		
Low	Mid	High	Low	Mid	High	Low	Mid	High

Figure 2: The alignment of the STAMP scale with the ACTFL proficiency scale. (From Santos, 2022, p. 2; reprinted with permission.)

Benchmark	Group	Reading	Listening
Advanced	High	617 – 660	622 – 671
	Mid	597 – 616	599 – 621
	Low	580 – 596	582 – 598
Intermediate	High	541 – 579	534 – 581
	Mid	523 – 540	521 – 533
	Low	502 – 522	501 – 520
Novice	High	461 – 501	466 – 500
	Mid	449 – 460	455 – 465
	Low	331 – 448	343 – 454

Figure 3: STAMP English reading and listening scaled scores in relation to the ACTFL proficiency scale. (From Avant Assessment, 2023; reprinted with permission.)

Novice (Low, Mid, High), Intermediate (Low, Mid, High), and Advanced (Low, Mid, High).

Figure 3 shows an interpretation of the scaled scores in relation to the ACTFL proficiency scale. Compared to ordinal proficiency ratings, the scaled scores provide a more fine-tuned, and thus more precise, view of a test taker's proficiency. For example, if a group of test takers all receive STAMP level 6 (Intermediate High) in listening, their corresponding scaled scores are within the range of 534–581, but can vary by up to 47 points. In other words, scaled scores further differentiate same-level test takers. For this reason, the main findings of this study are reported based on scaled scores.

3.3 Testing

An email soliciting participation was sent to a sample of Duolingo learners who pre-qualified for study participation in each course (based on prior proficiency and course completion criteria, as described in section 3.1 Participants). Learners interested in participating completed a background survey that allowed us to verify their eligibility and collect additional demographic information. Among the survey responders, those who reported that they

were younger than 18, had English as their home language before age 6, or had taken classes or used other apps/programs to learn English during the time they used Duolingo were disqualified from participation.

Participants were not guided or prompted in any way during the study period. The participants learned at their own pace and reached the end of the course section naturally. In other words, there was no control, guidance, or prompting from the authors, thereby maximizing the study's ecological validity.

Qualified participants were notified and invited to take the reading and listening sections of the STAMP 4S English Test. Data were collected during multiple test windows on a rolling basis, each lasting two weeks (from initial invitation to taking the test). Remote human proctors from Avant Assessment were present for each scheduled testing session. Each participant received US$75 and their score report after taking the test.

4. Results

As explained in section 3.2 (Instruments), the participants' reading and listening performances were evaluated using both STAMP levels and scaled scores. The STAMP levels are on an ordinal scale of 1–9, which corresponds to the ACTFL proficiency scale of Novice Low (1) to Advanced High (9). Each STAMP level corresponds to a range of scaled scores, which further differentiates same-level learners and provides a more precise understanding of their proficiency. Table 4 shows the number of participants who scored at each STAMP (and ACTFL) level across the scale in reading and listening.

Table 4:
Distribution of Scores at Each STAMP (and ACTFL) Level in Reading and Listening

			STAMP/ACTFL level								
			Novice			Intermediate			Advanced		
Course	*N*	Skill	Low 1	Mid 2	High 3	Low 4	Mid 5	High 6	Low 7	Mid 8	High 9
EN<JA	81	Reading	0	0	6	6	21	36	4	5	3
		Listening	0	0	6	18	11	41	3	2	0
EN<ES	72	Reading	0	0	3	5	2	32	15	8	7
		Listening	0	0	15	15	5	29	7	1	0
EN<PT	92	Reading	0	0	5	4	5	44	15	13	6
		Listening	0	0	6	21	14	43	4	3	1

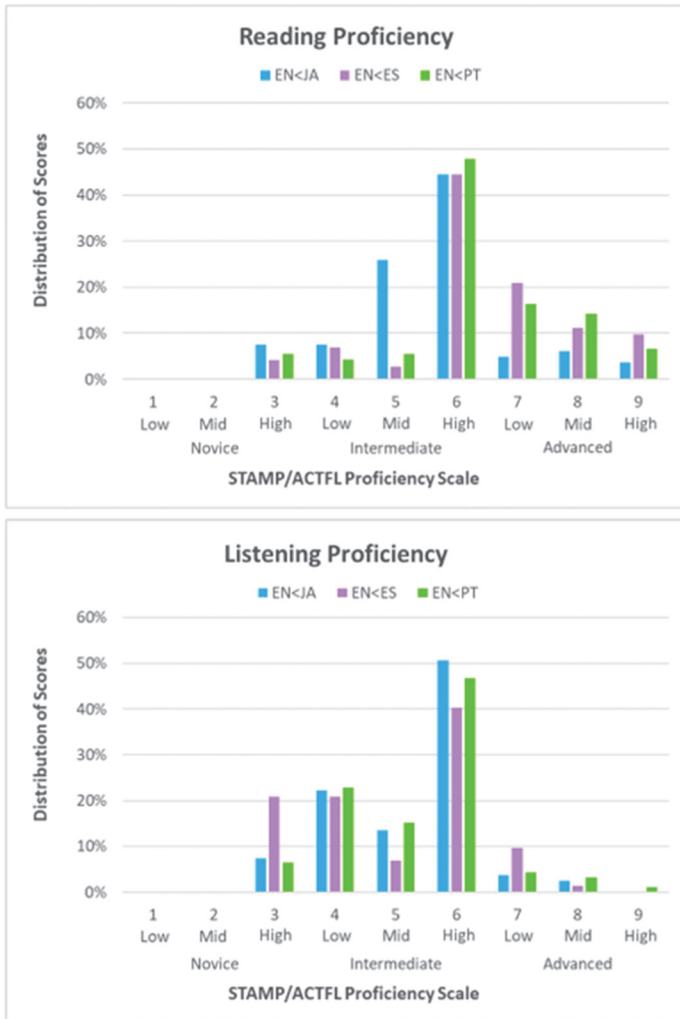

Figure 4: Reading and listening score distributions at each ACTFL level at the end of A2.

Figure 4 provides a visual presentation of score distributions in reading and listening.

The score distributions show that Intermediate High (level 6) is the mode of the distribution (i.e., the most frequent rating) for both reading and listening in all three courses. They also show a slight skew toward higher ratings in reading (except for the EN<JA course) and lower ratings in listening for all three courses.

Table 5:
Reading and Listening Proficiency in STAMP Scaled Scores at the End of A2

		Reading		Listening	
Course	***N***	**Mean scaled score (SD)**	**ACTFL sublevel**	**Mean scaled score (SD)**	**ACTFL sublevel**
EN<JA	81	550.33 (33.78)	Intermediate High	538.84 (28.10)	Intermediate High
EN<ES	72	568.54 (36.32)	Intermediate High	543.51 (35.85)	Intermediate High
EN<PT	92	567.49 (33.04)	Intermediate High	539.09 (30.48)	Intermediate High

Table 5 shows the average scaled scores of the participants and their alignment with the ACTFL scale. At the completion of A2 in the course, the Japanese-speaking participants averaged 550 in reading and 539 in listening; the Spanish-speaking participants averaged 569 in reading and 544 in listening; and the Portuguese-speaking participants averaged 567 in reading and 539 in listening. Since a different pattern of distribution was noticed in the reading score of the Japanese group and their average scaled score in reading was lower compared to the other two groups, we ran an analysis of variance to evaluate whether the differences reached statistical significance. We found that the reading score of the Japanese group was significantly lower than that of the Spanish or Portuguese group, but no differences were found in listening scores among the three groups. No statistically significant differences were found between the Spanish and the Portuguese groups in either reading or listening scores. Importantly, all these average scaled scores were within the range of Intermediate High for both reading and listening (see Figure 3). These results are also presented graphically in Figure 5.

5. Discussion

This study evaluated learners upon completion of the Basic content (through CEFR A2) of three Duolingo English courses: English for Japanese, Spanish, and Portuguese speakers. On average, the participants scored Intermediate High in both reading and listening based on STAMP scaled scores. The consistent and consolidated evidence from all three courses strengthens the evidence that Duolingo's Basic content in English courses is effective in developing learners' reading and listening skills.

This study involved two different proficiency frameworks in the field of language learning and assessment: the CEFR and the ACTFL guidelines. As

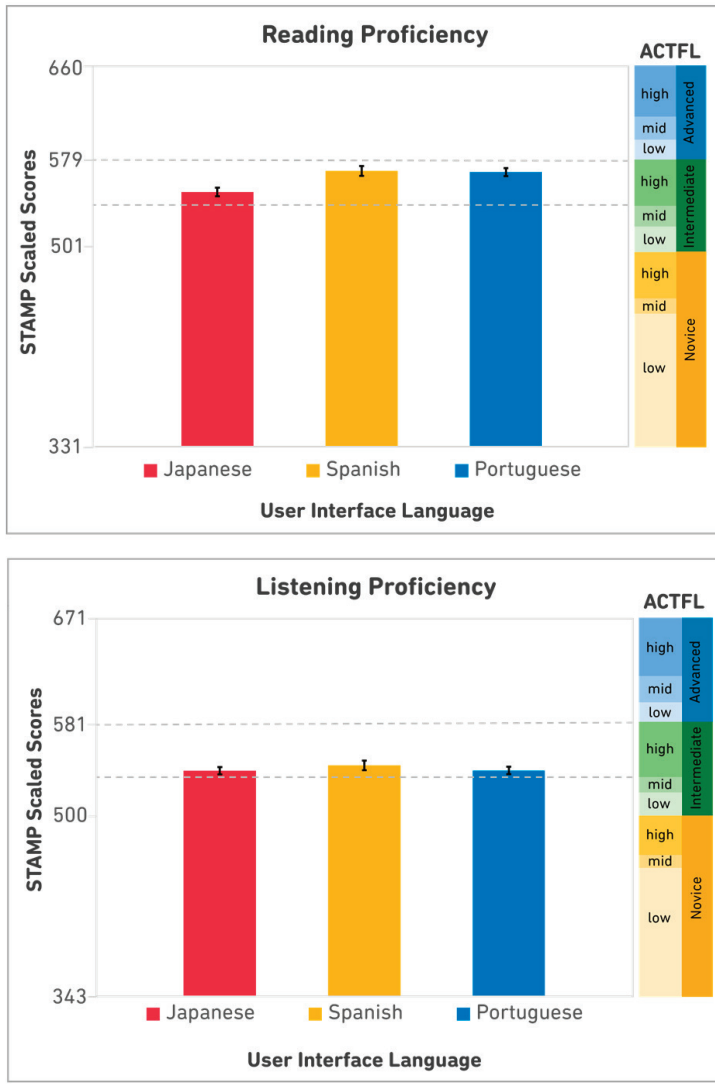

Figure 5: Reading and listening proficiency of participants at the end of A2, with Standard Error of Measurement (SEM).

well-established educational standards, they both provide the basis for curriculum development, test development, and test score interpretation. This study involved both frameworks. The Duolingo English courses were aligned with the CEFR, and the participants were assessed when they completed the CEFR A2 content, while the interpretation of their test scores was based on the ACTFL proficiency guidelines.

Although the two frameworks have co-existed for more than 20 years, few empirical studies have investigated the correspondence between them. Currently, the interpretation of the STAMP 4S English Test scores is only aligned to the ACTFL proficiency guidelines, and Avant Assessment does not provide a concordance between their STAMP scale and the CEFR. However, Avant Assessment (Santos, May 22, 2023, personal communication) expects that "a STAMP level 6 (Intermediate High) will map to a CEFR B1, given their current work on developing STAMP 4S CEFR tests and previous research in this area."

ACTFL (n.d.) published an empirically based alignment between the two frameworks and uses it to assign CEFR levels to their own assessments. For the ACTFL reading and listening proficiency tests, CEFR A2 corresponds to ACTFL Intermediate Mid, and CEFR B1 is aligned to ACTFL Intermediate High (B1.1) and Advanced Low (B1.2). Based on these correspondences, the reading and listening proficiency of the participants in the current study is one ACTFL sublevel above our expected course outcomes of Intermediate Mid. In other words, the participants of this study scored above A2 and at early B1 after they completed the A2 content on Duolingo's English courses for Japanese, Spanish, or Portuguese speakers, which, to some extent, demonstrates the effectiveness of these courses in developing learners' reading and listening skills.

Nevertheless, the absence of a pretest raises a plausible concern that some participants could potentially be false beginners, or that they may have underestimated their initial proficiency level when they started learning on Duolingo. To mitigate these concerns, additional data on the participants' self-perceived proficiency has been incorporated and displayed in Table 6. Participants were requested to evaluate their English proficiency on a scale ranging from 0 to 10, where 0 signifies "no knowledge" and 10 represents "fluent." This was part of the background survey that the participants completed during data collection. The questions sought their perception at two different points: upon completion of the A2 content and at the outset of their English learning journey on Duolingo. The specific questions were: (1) "How much English do you think

Table 6:
Perceived Proficiency Gains

Course	*N*	Before learning on Duolingo (0–10)	After finishing A2 on Duolingo (0–10)	Perceived gain
EN<JA	81	3.21	5.09	1.88
EN<ES	71	1.58	5.28	3.70
EN<PT	92	1.59	5.46	3.87

you knew before studying on Duolingo?" and (2) "How much English do you think you know now?" The perceived proficiency gains were computed by determining the difference between the responses to question 2 and question 1.

As mentioned in section 3.1 (Participants), we assessed the Japanese-speaking learners in the EN<JA course after the Duolingo homescreen transition. Together with the homescreen transition, the prior proficiency scale Duolingo uses when onboarding new users also changed from a 0–10 scale to a four-level scale. Most participants in the EN<JA course indicated that they were familiar with "basic words and phrases," with the exception of two participants who were new to English. Based on their perceived proficiency before starting with Duolingo (Table 6), they were not exactly true beginners (with an average rating of 3.21 out of 10). This is also supported by their initial placement in the Duolingo course (as shown in Table 3). Only over a third of these participants started at the beginning section of the course, while more than 80% of the Spanish- or Portuguese-speaking learners started from the very beginning of their course. Therefore, having higher prior knowledge in this case means going through less content in the course. Despite having a bit more prior knowledge than learners in the other two courses, the EN<JA course participants did not show higher proficiency scores; in fact, their reading proficiency was lower than the Spanish and the Portuguese groups by the time they completed A2.

Language learning apps have been criticized as only being effective for acquiring vocabulary and grammar in a decontextualized setting (Krashen, 2014). Despite Duolingo's initial lessons focusing primarily on sentence-level vocabulary and grammar, this study shows that users can successfully apply their learned linguistic knowledge to integrative tasks like reading and listening comprehension. The application and integration of linguistic knowledge was also noted in Loewen et al. (2020), where Babbel users showed promising improvements in speaking proficiency even with limited practice of oral language. Skill acquisition theory, as elaborated by DeKeyser (2015), supports this phenomenon, suggesting that repeated practice can transition explicit knowledge into procedural knowledge, enhancing language learning outcomes. Reflecting on these findings, Loewen et al. (2020, p. 19) suggested that the field of SLA should re-evaluate its views on language learning apps, moving away from seeing them as simple grammar drill machines, and recognize their pedagogical potential, a view supported by Heift and Vyatkina (2017). The authors of this study align with Loewen et al.'s suggestion. However, any assertions about skill transfer among Duolingo users should be empirically verified.

6. Limitations and Future Direction

Although the design of the study presents some level of ecological validity because the participants reached the end of A2 naturally and independently, future research will benefit from more controlled designs such as a pre- and posttest design and/or a comparison-group design, neither of which was planned for the present study due to logistical and access-related constraints, but which we intend to integrate into future studies. These design features will allow more control and/or measurement of learning time, as well as other participant factors that were self-reported in the present study. Future research examining users' in-app experience and learners' perceptions thereof would also be useful as a means to document and improve our understanding of app-based language learning from the learners' perspective.

Because the three English courses were assessed one at a time over a 14-month period and the Duolingo product changes often, there were some differences in what participants experienced in each course. One of the changes was the shift from a 0–10 rating scale to a four-level rating scale on prior proficiency. The Japanese-speaking learners' placement data in Table 3 seemed to indicate that they were not true beginners, even after we limited participation to those whose self-reported prior proficiency was level 2 on the four-level scale—only knowing basic words and phrases. Learners at that level seemed to know more than those whose prior proficiency was 0–2 on the 0–10 rating scale. For future research in this line of inquiry, it might be useful to limit study participation to those who report prior proficiency at level 1 on the four-level scale.

We would also note that the skills of reading and listening assessed in the study are both receptive. Learners were not assessed in productive skills such as speaking and writing, or overall proficiency. Future studies should evaluate Duolingo's effectiveness in developing English learners' productive skills or overall proficiency. Doing so will lead to a better understanding of whether and to what extent Duolingo English learners' success in receptive skills can also be observed in productive skills or overall proficiency.

7. Conclusion

In sum, this study evaluated the reading and listening proficiency outcomes of Duolingo English learners who self-reported having little to no prior knowledge of the target language and used Duolingo as their only learning tool. Evidence from three English courses demonstrated that participants who completed the Basic content (through CEFR A2) reached, on average,

Intermediate High in both reading and listening. These proficiency scores indicate that the Duolingo English courses for Japanese, Spanish, and Portuguese speakers are effective in developing learners' reading and listening skills.

Note

1. The Duolingo placement test is an integrated feature that enables users to begin a new course at a more advanced level. This adaptive test selects questions dynamically from various course sections. As the user answers correctly, the test becomes increasingly difficult, and vice versa when they answer incorrectly. Although the test comprises 20 questions, it will terminate prematurely if the user answers 7 consecutive questions incorrectly.

Acknowledgments

We would like to thank Audrey Kittredge, Elise Hopman, Erin Gustafson, and Lucy Skidmore for their help at different stages of the project.

About the Authors

Xiangying Jiang is a lead learning scientist at Duolingo. She works on learning assessment and efficacy research at Duolingo. She has a PhD in Applied Linguistics (Northern Arizona University, 2007). Prior to joining Duolingo in 2019, she was a faculty member at West Virginia University. Her work has appeared in journals such as *The Modern Language Journal*, *Reading in a Foreign Language*, *Reading Psychology*, and *Foreign Language Annals*.

Ryan Peters is a learning scientist at Duolingo. He has a PhD in Speech, Language and Hearing Sciences (Purdue University, 2019). Prior to joining Duolingo in 2022, he held dual positions as a Postdoctoral Researcher in the Department of Psychology at the University of Texas at Austin and as a Museum Researcher at Thinkery, a children's science museum in Austin. His research focused primarily on the interplay between attention and learning, with an emphasis on first and second language learning.

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

Bozena Pajak holds a PhD in Linguistics (University of California, San Diego, 2012). Before joining Duolingo in 2015, she was a Research Associate and a Lecturer in Linguistics at Northwestern University, Illinois. Her research focused primarily on the acquisition of additional languages in adulthood. She is currently the Vice-President of Learning and Curriculum at Duolingo.

References

ACTFL. (2012). *ACTFL proficiency guidelines*. Retrieved June 2, 2023 from https://www.actfl.org/resources/actfl-proficiency-guidelines-2012

ACTFL. (n.d.). *Assigning CEFR ratings to ACTFL assessments*. Retrieved June 2, 2023 from https://www.actfl.org/uploads/files/general/Assigning_CEFR_Ratings_To_ACTFL_Assessments.pdf

Ajisoko, P. (2020). The use of Duolingo apps to improve English vocabulary learning. *International Journal of Emerging Technologies in Learning*, *15*(7), 149–155. https://doi.org/10.3991/ijet.v15i07.13229

Ajisoko, P. (2022). Using Duolingo apps to improve English reading comprehension of engineering students in Universitas Borneo Tarakan. *Exposure: Jurnal Pendidikan Bahasa Inggris* [English Language Education Journal], *11*(1), 1–6. https://doi.org/10.26618/exposure.v11i1.6452

Avant Assessment. (2023). *STAMP scaled scores guide*. Retrieved June 2, 2023 from https://avantassessment.com/stamp-scaled-scores-guide

Brown, K. (2023). *CALL as a driver of equity, access, and outcomes in an increasingly connected world*. Opening plenary at CALICO Annual Conference, June 7, 2003, University of Minnesota, Minneapolis.

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

Bustillo, J., Rivera, C., Guzman, J. G., & Acosta, L. R. (2017). Benefits of using a mobile application in learning a foreign language. *Sistemas Y Telematica*, *15*(40), 55–68. https://doi.org/10.18046/syt.v15i40.2391

Carpenter, S. K. (2020). Distributed practice/spacing effect. In L. Zhang (Ed.), *Oxford research encyclopedia of education* (pp. 1–20). New York: Oxford University Press. https://doi.org/10.1093/acrefore/9780190264093.013.859

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

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

Fatmawati, I., Sudirman, A., & Munawaroh, T. (2023). The effect of using Duolingo application towards students' grammar mastery in simple present tense. *Journal of English Pedagogy and Applied Linguistics*, *3*(2), 13–21. https://doi.org/10.32627/jepal.v3i2.576

Heift, T., & Vyatkina, N. (2017). Technologies for teaching and learning L2 grammar. In C. Chapelle & S. Sauro (Eds.), *The handbook of technology and second language teaching and learning* (pp. 26–44). New York: John Wiley & Sons. https://doi.org/10.1002/9781118914069.ch3

James, K. K., & Mayer, R. E. (2019). Learning a second language by playing a game. *Applied Cognitive Psychology*, *33*(4), 669–674. https://doi.org/10.1002/acp.3492

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

Kessler, M., Loewen, S., & Gonulal, T. (2023). Mobile-assisted language learning with Babbel and Duolingo: Comparing L2 learning gains and user experience. *Computer Assisted Language Learning*, *36*. https://doi.org/10.1080/09588221.2023.2215294

Kim, S. K., & Webb, S. (2022). The effects of spaced practice on second language learning: A meta-analysis. *Language Learning*, *72*(1), 269–319. https://doi.org/10.1111/lang.12479

Krashen, S. (2014). Does Duolingo "trump" university-level language learning? *International Journal of Foreign Language Teaching*, 13–15. Retrieved June 2, 2023 from https://sdkrashen.com/content/articles/krashen-does-duolingo-trump.pdf

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

Loewen, S., & Hui, B. (2021). Small samples in instructed second language acquisition research. *Modern Language Journal*, *105*(1), 187–193. https://doi.org/10.1111/modl.12700

Loewen, S., Crowther, D., Isbell, D., Kim, K., Maloney, J., Miller, Z., & Rawal, H. (2019). Mobile-assisted language learning: A Duolingo case study. *ReCALL*, *31*(3), 293–311. https://doi.org/10.1017/S0958344019000065

Loewen, S., Isbell, D. R., & Sporn, Z. (2020). The effectiveness of app-based language instruction for developing receptive linguistic knowledge and oral communicative ability. *Foreign Language Annals*, *53*(2), 209–233. https://doi.org/10.1111/flan.12454

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

Norouzian, R. (2020). Sample size planning in quantitative L2 research: A pragmatic approach. *Studies in Second Language Acquisition*, *42*(4), 849–870. https://doi.org/10.1017/S0272263120000017

Norris, J. M., & Ortega, L. (2012). Assessing learner knowledge. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 573–589). New York: Routledge.

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

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, *35*(4), 655–687. https://doi.org/10.1017/S0272263113000399

Plonsky, L. (2017). Quantitative research methods. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 505–521). New York: Routledge. https://doi.org/10.4324/9781315676968-28

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

Rohrer, D. (2015). Student instruction should be distributed over long time periods. *Educational Psychology Review*, *27*(4), 635–643. https://doi.org/10.1007/s10648-015-9332-4

Rolando, A. P. P., Gabriela, R. E. A., Carolina, M. S. E., & Antonio, A. R. M. (2019). Incidencia de Duolingo en el desarrollo de las habilidades comunicacionales verbales del

idioma ingles a nivel de educación superior. *European Scientific Journal*, *15*(16), 29–44. https://doi.org/10.19044/esj.2019.v15n16p29

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

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

Sudina, E., & Plonsky, L. (2023). The effects of frequency, duration, and intensity on L2 learning through Duolingo: A natural experiment. *Journal of Second Language Studies, 7*(1), 1–43. https://doi.org/10.1075/jsls.00021.plo

Tarone, E. (Ed.) (2015). Online foreign language education: What are the proficiency outcomes? *Modern Language Journal*, *99*(2), 392–415. https://doi.org/10.1111/modl.12220

Thomas, M. (2006). Research synthesis and historiography: The case of assessment of second language proficiency. In L. Ortega & J. M. Norris (Eds.), *Synthesizing research on language learning and teaching* (pp. 279–298). Philadelphia: John Benjamins. https://doi.org/10.1075/lllt.13.13tho

Winke, P., Gass, S. M., Soneson, D., Rubio, F., & Hacking, J. F. (2020). *Foreign language proficiency test data from three American universities [United States], 2014–2017*. Inter-University Consortium for Political and Social Research.

Appendix: Characteristics of the Study Participants

Characteristics	EN<JA (*N* = 81)	EN<ES (*N* = 72)	EN<PT (*N* = 92)
Age			
18–34 years	34.46%	38.89%	42.39%
35–54 years	64.20%	47.22%	38.04%
55–74 years	12.35%	13.89%	19.57%
Home language before age 6			
Only Japanese/Spanish/ Portuguese	100%	93.06%	92.39%
Language other than Japanese/Spanish/ Portuguese	0%	4.17%	4.35%
More than one language	0%	2.78%	3.26%
Highest level of education			
Bachelor's degree	53.09%	44.44%	41.30%
Master's degree	13.58%	19.44%	8.70%
Doctoral degree	0%	4.17%	3.26%
High school	16.05%	5.56%	4.35%
Technical/professional	12.34%	20.83%	31.52%
Other	4.94%	5.56%	10.87%
Primary reason for learning English			
Job-related purposes	39.51%	54.17%	53.26%
School	28.40%	47.22%	53.26%
Travel	19.75%	41.67%	46.74%
Social purposes	39.51%	41.67%	34.78%
Memory/brain acuteness	17.28%	37.50%	34.78%
Fun/leisure	34.57%	26.39%	29.35%

Note: Learners could choose several options as their primary reason for learning English. As a result, the percentages do not add up to 100%.