

The effects of frequency, duration, and intensity on L2 learning through Duolingo

A natural experiment

Ekaterina Sudina and Luke Plonsky

East Carolina University, United States | Northern Arizona University,
United States

Instructed second language (L2) research has frequently addressed the effects of spacing, or, alternatively, the distribution of practice effects. The present study addresses Rogers and Cheung's (2021) concerns about the ecological validity of such work via a natural experiment (Craig et al., 2017). Learners' self-determined exposure and in-app behavior were examined in relation to language gains over time. Duolingo learners of Spanish or French ($N=287$) completed a background questionnaire, scales measuring L2 motivation and grit, and two tests of L2 proficiency before and after a six-month period of user-controlled app usage. Total minutes of app exposure exhibited a correlation with written but not oral proficiency gains. More dependable correlates of gains were frequency- and curriculum-oriented measures. Additionally, L2 grit and motivation were weakly to moderately correlated with several in-app behaviors. We conclude with implications for how apps can best be leveraged to produce L2 gains.

Keywords: distribution of practice, instructed SLA, mobile-assisted language learning, grit, Duolingo

1. Introduction

Second-language (L2) development is greatly influenced by the type and amount of instruction that learners receive (e.g., Norris & Ortega, 2000; Saito & Plonsky, 2019). Less clear, however, are the effects of frequency, duration, and intensity of target language exposure and practice. That is, are intensive periods of instruction most effective, or is instruction more effective through more frequent exposure (see Carpenter, 2020; Rogers & Cheung, 2021)?

Informed by a quickly growing body of research in second-language acquisition (SLA) that has sought to address this question (e.g., Kasprowicz et al., 2019; Li & DeKeyser, 2019; Yamagata et al., 2023), the present study builds on these and other recent studies of distributed learning in SLA (e.g., Serrano & Huang, 2018; Suzuki & DeKeyser, 2017) as well as several decades of theoretical and empirical attention in educational psychology (see Carpenter, 2020; Rohrer, 2015) to shed light on the effects of frequency and duration of Duolingo app usage on L2 learning in the context of the distribution of practice. In doing so, the study seeks to make several unique and worthwhile contributions.

First, being the first to examine distributional effects for app-based language learning, this study allows us to gain a better understanding of the effects of frequency, duration, and intensity specifically for the Duolingo app and its users. Second, previous studies on distributed learning have targeted predominantly vocabulary and grammar knowledge in either traditional classroom or laboratory settings (e.g., Kasprowicz et al., 2019; Nakata & Suzuki, 2019; Suzuki, 2019). The present study, by contrast, considers overall proficiency in both oral and written modes in a new and flourishing context of instructional technology. Third, Rogers and Cheung (2021) raised concerns over the ecological validity of previous work on distributed learning. In particular, the authors questioned whether existing results, which are almost exclusively obtained in labs, would hold in less controlled settings. The present study addresses this concern by conducting a natural experiment (Craig et al., 2017), whereby learners' self-determined exposure and behavior (i.e., frequency, duration, and intensity of app usage) are examined in relation to language proficiency gains made over a six-month period. Fourth, in addition to exposure effects, two individual differences, L2 grit (Teimouri et al., 2022) and motivation (Papi et al., 2019), are modeled in relation to learner behavior and gains in proficiency to isolate and better understand the effects of frequency, duration, and intensity of instruction. The literature review that follows offers an overview of theories and studies on distributed learning in SLA and neighboring disciplines, introduces the concept of natural experiment, surveys recent studies on the role of app-based technology in instructed SLA, and discusses the concepts of L2 grit and motivation that are relevant to the present study. The last section of the literature review provides an overview of the Duolingo course structure at the time of data collection to better situate the study.

2. Literature review

2.1 Distributed learning

Rogers and Cheung (2021) provided definitions of several concepts pertinent to the present investigation. First, “*distribution of practice*, also referred to as *input spacing*, refers to whether and how learning is spaced over multiple learning episodes” (pp.1138–1139; see also Rogers, 2017 for an overview of theoretical and methodological differences in research on the distribution of practice in SLA and cognitive psychology). Second, “*massed practice* refers to experimental conditions in which learning is concentrated into a single, uninterrupted training session, whereas *distributed* or *spaced practice* refers to learning that is spread over two or more training episodes” (p.1139). It is also important to differentiate between spacing and lag effects as the latter were the focus of the current study. Spacing effects (i.e., comparison of a massed condition versus a spaced condition) are typically robust; lag effects (i.e., advantage of a longer gap over a shorter gap) are harder to obtain (see Rogers’ 2023 conceptual review for more); data from Kim and Webb’s (2022) meta-analysis support this finding as an advantage of longer spacing was found in delayed but not in immediate posttests. As noted by Rogers (2023), “spaced-condition groups outperform the massed-condition group with relatively large effect sizes ($d \geq 1.0$), but there is little to no difference in performance between those in the different distributed conditions” (p.452).

According to Carpenter (2020), there is a consensus in educational research that distributed practice, or spacing effect, is more conducive to learning when repeated instruction is dispersed over time rather than happens in rapid succession (for a review of spacing and massing, see Rohrer, 2015). This stems in part from deficient processing theory, which postulates that spaced instruction provides learners with more opportunities for noticing and ultimately leads to better retention of the information presented. Another theory that might explain the spacing effect is that of encoding variability; it emphasizes the role of contextual cues, which become more pronounced in spaced learning experiences. Next, a study-phase retrieval theory suggests that retrieval of learners’ previous experience, which typically happens during distributed instruction, aids in future retention of the material. Finally, consolidation in the form of neural activation processes is more likely to occur during spaced rather than massed practice. Critically, although the aforementioned theories have made great strides in explaining spacing effects, none of them is fully satisfactory as a stand-alone theory; thus, there is currently no consensus on the underlying theoretical mechanism responsible for spacing and lag effects in learning (see Delaney et al., 2010, for more).

It should be noted that spaced instructional activities may be challenging to implement in real-life classrooms as this approach requires considerable planning on the part of instructors who need to create multiple lesson plans in advance and adhere to spacing schedules consistently for longer periods of time. One could even make the case that spacing effects are not as relevant in an ecological sense to L2 classroom practice, or to mobile-assisted language learning (MALL). SLA researchers are typically interested in lag effects, and comparisons of shorter gaps (e.g., a 1-day gap between the time that the learner logs in to practice) versus a longer gap (e.g., a 7-day gap). Notably, despite the consensus towards spacing effects across the broader psychological literature (as suggested by Carpenter, 2020), there are fewer studies (and a lack of consensus) on lag effects (e.g., Rohrer, 2015; Rogers & Cheung, 2020).

Moreover, according to Cepeda et al. (2008), a true spacing study from a cognitive psychology viewpoint includes “multiple periods of study devoted to the same material, separated by some variable time gap, with a final memory test administered after an additional retention interval” (p.1095). Here, the “same material” is an important feature of a spacing experiment. Thus, those SLA studies that do not include the exact same content are arguably not truly spacing studies. This questions the generalizability of cognitive psychology findings to instructed SLA (and by extension MALL) where the language content is rarely copied verbatim from one session to the next.

Furthermore, Rogers and Cheung (2021) contended that experimental research in the domain of distributed learning tends to overemphasize internal validity of a study to the detriment of its external and ecological validity, which limits the generalizability of tightly controlled experiments to authentic instructional settings. In fact, the enhanced ecological validity of Rogers and Cheung’s (2021) conceptual replication was arguably one of the reasons why their study did not lend support to the advantages of a more distributed L2 vocabulary learning practice (i.e., 8 days between training sessions) compared to a less distributed one (i.e., 1 day between training sessions) among English-as-a-foreign language (EFL) child participants in Hong Kong. This finding largely contradicted previous lab-based research with adult participants studying L2 vocabulary (e.g., Nakata & Suzuki, 2019). However, Rogers and Cheung’s (2021) results resembled those of Kasprovicz et al. (2019) – another study with enhanced ecological validity that examined the effects of distributed practice among young learners of French studying L2 grammar (verb morphology). The interval being examined by Kasprovicz et al. (2019), although quite constrained (3.5 vs. 7 days), emulated “the most common lesson frequency in UK primary schools (one or two lessons per week)” (p.585); yet it did not yield differences in learning between groups. By contrast, Li and DeKeyser (2019) found enhanced procedural knowledge reten-

tion among adult learners of Mandarin Chinese exposed to instruction with shorter rather than longer intervals between sessions (1 day vs. 1 week), but these results did not hold for declarative knowledge. Similarly complex results were reported by Serrano and Huang (2018). In their study, intense repeated L2 reading sessions (1-day interval) were more beneficial for short-term gains in L2 vocabulary for teenage EFL learners in Taiwan (following the immediate posttest), whereas spaced sessions (1-week interval) resulted in higher retention long-term (from the immediate to the delayed posttest). Nonetheless, the difference between the two groups was negligible when vocabulary gains were compared from the pretest to the delayed posttest.

To make the picture even more complex, Suzuki and DeKeyser's (2017) study of L2 Japanese morphology in adult learners found no advantages of distributed practice (7-day interval) over massed practice (1-day interval) for utterance *accuracy*; moreover, it was less conducive to utterance *fluency* than massed practice. However, the authors emphasized that the results of their lab-based study may not be generalizable to real-life classrooms. Notably, Kim and Webb's (2022) meta-analysis of distributed practice revealed the advantage of spaced learning in SLA, as indicated by small-to-medium and medium-to-large effect sizes (i.e., $g=0.58$ and 0.80 for L2 learning and retention, respectively) across 37 eligible studies (48 independent experiments) in their sample. Critically, the researchers argued that the spacing effect varied depending on the methodological features of the study design, which explains in part the complex and conflicting results of the primary studies reviewed above. Moreover, the increasing maturity of spacing research within SLA is evidenced in replication studies. To illustrate, Suzuki and DeKeyser's (2017) complex findings have been replicated by Suzuki (2017). Other examples include Serrano and Huang (2018, 2023) as well as Rogers and Cheung (2020, 2021). For an overview of a broader spacing literature, see Serrano (2022).¹

2.2 Natural experiment

Despite the comprehensiveness of Kim and Webb's (2022) meta-analysis, one methodological variable that was not examined as a moderator in their sample was that of ecological validity of primary studies on distributed spacing. Apart from experimental designs, high ecological validity bears relevance for observational designs as well. One example of an observational study with increased ecological validity is a natural experiment, which refers to "any event not under the control of a researcher that divides a population into exposed and unexposed

1. We thank an anonymous reviewer for their considerable input into this section of the Literature review.

groups”; this allows natural experiments to “use this naturally occurring variation in exposure to identify the impact of the event on some outcome of interest” (Craig et al., 2017, p.2). Critically, one of the major differences between randomized controlled trials, natural experiments, and nonexperimental observational studies is how well the intervention has been defined (Craig et al., 2017, Table 1). In randomized controlled trials, the intervention is thoroughly documented and implemented; in natural experiments, the intervention is also happening, but researchers have less control over it; finally, in nonexperimental observational studies, there is no clear intervention at all. Although natural experiments have been particularly embraced in public health research, they have immediate applicability in instructed SLA and applied linguistics research more generally, especially in situations where experimental manipulations are not deemed reasonable or ethical. As such, natural experiments have arguably higher ecological validity compared to more traditional, rigidly controlled experiments. One example of a natural experiment for SLA would be a study that examines the effect of participating in a study abroad program on L2 learners’ oral proficiency using a pre- and posttest design (unlike a quasi-experimental study, a natural experiment does not strictly control the independent variables). However, to our knowledge, no study to date has conducted a natural experiment to investigate the effectiveness of mobile-assisted language learning.

2.3 Mobile-assisted language learning

As noted by Loewen (2020), the use of technology is now considered to be one of four major contexts in instructed SLA along with traditional classroom, study abroad, and immersion instruction (see Loewen, 2020). One of the advantages of instructional technology is that it has “the potential to speed up or enhance the process” of L2 learning and, most importantly, “deliver individualized instructional materials that meet learners at their specific levels of proficiency” (Loewen, 2020, p.193). A second-order synthesis by Plonsky and Ziegler (2016) found a small advantage of computer-assisted language learning over traditional classroom instruction (based on the results of 14 meta-analyses in this domain, which included a total of 408 primary studies and over 14,000 language learners). Nonetheless, this synthesis was unable to examine the effectiveness of mobile-assisted language learning due to the lack of primary studies in this area.

The situation, however, is rapidly changing, and research focusing on the role of mobile- and app-based technology in instructed SLA is currently on the rise (e.g., García Botero et al., 2019; Jiang, Rollinson, et al., 2021; Loewen et al., 2019, 2020). In fact, several recent studies have focused specifically on the relative effectiveness of Duolingo vs. university-based language instruction. The beginner

and intermediate Duolingo learners attained similar – and in some cases, superior – L2 proficiency levels as university students who studied foreign languages for four and five semesters, respectively (see Jiang et al., 2020; Jiang, Chen, et al., 2021). Of particular relevance to the present investigation is a study by Jiang, Rollinson, Plonsky, et al. (2021) that explored a possible relationship between app usage and gains in proficiency and observed modest correlations between the two ($\rho = .02-.14$ for L2 French listening and reading; $\rho = .01-.06$ for L2 Spanish listening and reading, respectively). However, only one of many possible temporal or exposure-related indicators was used (total hours). Furthermore, the sample included only novices, and no pretest data were collected – features the present study seeks to improve on.

Additionally, recently published studies and existing reports on Duolingo effectiveness have predominantly assessed specific language skills (i.e., listening, reading, and speaking; see Jiang et al., 2020; Jiang, Chen, et al., 2021; Jiang, Rollinson, et al., 2021; Jiang, Rollinson, Plonsky, et al., 2021).² One exception is Loewen et al.'s (2019) investigation of Duolingo users' overall L2 Turkish proficiency along with five subareas (listening, speaking, writing, reading, and lexicogrammar); although comprehensive and informative, this study did not involve a control or comparison group and included only nine participants. Clearly, the field stands to benefit from more research on L2 learners' overall proficiency in the domain of mobile-assisted language learning.

2.4 L2 grit and motivation

Although research on individual differences has long established its niche in SLA overall (Gass et al., 2020) and computer-assisted language learning in particular (see Pawlak & Kruk, 2022), much remains uncertain about the role of individual differences in mobile-assisted language learning. For example, Loewen et al. (2019) raised concerns over participants' attitudes to some app-related features (e.g., lack of interaction and limited variation in Duolingo tasks), which might have affected learners' motivation and persistence in language app use. In the same vein, García Botero et al. (2019) noted inconsistencies between Duolingo learners' questionnaire responses, which pointed to students' motivation in and positive attitude towards using the app out of class, and their interview data, which demonstrated students' mixed views on engagement and lack of long-term interest while using the app.

2. The first three publications are white papers published by Duolingo; Jiang, Rollinson, Plonsky, et al., 2021 is a peer-reviewed article.

To investigate these issues further, research into mobile-assisted language learning would benefit from examining language app users' academic perseverance, or grit, as well as their motivated learning behavior. Grit has been defined as "perseverance and passion for long-term goals" (Duckworth et al., 2007, p.1087); some researchers have also conceptualized it as a facet of the personality trait of conscientiousness (Park et al., 2018; Schmidt et al., 2018). Nonetheless, it becomes increasingly common in SLA research to conceptualize and measure grit as a domain-specific construct by tailoring scale items and instructions to a specific language learning context (see Teimouri et al., 2021 for more). Indeed, a growing body of research into L2 grit has found evidence of a positive relationship between language-domain-specific grit and achievement (e.g., Sudina & Plonsky, 2021a; Teimouri et al., 2022).

One of the constructs that is conceptually related to L2 grit is intended effort, defined as "the amount of time, effort, and energy L2 learners expend in the process of L2 learning" (Teimouri, 2017, p.686) and recently reconceptualized into current L2 motivated learning behavior by Papi et al. (2019). This was done in order to avoid bias in favor of promotion-focus rather than prevention-focus learners and tap into learners' actual motivational behavior rather than their hypothetical disposition (see Papi et al., 2019).

2.5 The Duolingo course structure

All Duolingo courses are aligned with the Common European Framework of Reference (CEFR). Both French and Spanish courses start with a brief Intro section (also known as A1.0, where A corresponds to the beginner or Basic User level; see Council of Europe, 2001) followed by the A1 content, which has two sections (A1.1 and A1.2) and covers both communicatively functional as well as grammatical topics, and the A2 content, which also consists of two sections (A2.1 and A2.2) and covers more advanced vocabulary and grammar. The last section of each Duolingo course includes B1 content, where B corresponds to the intermediate or Independent User level; see Council of Europe, 2001). The B1 content has four sections (B1.1 through B1.4), at the end of which language learners are expected to have mastered even more advanced communicatively functional and grammatical topics (e.g., "World news," "Learning," "Subjunctive with common conjunctions," and "Past conditional" for French; "World news," "Gossip," "Imperfect subjunctive," and "Passive" for Spanish). In addition to allowing for multiple "opportunities for practice and repeated exposure to target language structures," the Duolingo courses combine "more implicit, comprehension-based learning with explicit feedback and explanations" (Jiang, Rollinson, Plonsky, et al., 2021, p.981). Notably, Duolingo encourages a high "degree of user autonomy in navigating the

platform,” which translates into “substantial variation among individual learners on both the percentage of content they complete before reaching the end” of the B1.4 level and “on the total amount of time spent learning” (Jiang, Chen, et al., 2021, p. 2).

2.6 The present study

Expanding on previous research on distributed learning in SLA, the present study in the form of a natural experiment addressed the following research questions (RQ) concerning Duolingo effectiveness and L2 development more generally:

- RQ1. To what extent do learner gains differ when tested in the written vs. oral mode?
- RQ2. To what extent are frequency, duration, and intensity of Duolingo app usage associated with gains in L2 Spanish and French?
- RQ3. To what extent are L2 grit and motivation associated with the frequency, duration, and intensity with which learners use Duolingo?
- RQ4. To what extent are L2 grit and motivation associated with gains in L2 Spanish and French?

The present study meets the criteria for a natural experiment due to (a) “a clearly identified intervention” which was not rigidly controlled as the goal was to examine the effects of learners’ self-determined exposure and in-app behavior, (b) a lack of random assignment to intervention, and (c) a pre-posttest study design (Craig et al., 2017, p. 19).

3. Method

3.1 Participants

In the Fall of 2021, a group of 787 participants studying Spanish ($k=406$) or French ($k=381$) on Duolingo were invited to participate and completed a pretest (completion rate=34%). They were recruited at the beginning of the A1.2 section among beginner-level learners (see *The Duolingo Course Structure* in the Literature Review section). More specifically, the participants were at Row 18 of the French course tree structure (out of a total of 106 rows) and Row 21 of the Spanish course tree structure (out of a total of 121 rows), respectively. This suggests that the Duolingo course participants were already familiar with the basics (e.g., family and travel-related vocabulary and expressions, the present tense) as well

as slightly more advanced topics (e.g., shopping and routines-related vocabulary, grammatical agreement).

Six months later, in the Spring of 2022, a group of 288 participants completed the posttest (out of a total of 427 of those who met the selection criteria and eligibility requirements, see *Procedure*; response rate = 67%). One participant was excluded due to completing a pretest in Spanish and a posttest in French. Therefore, the final sample comprised 287 participants (Spanish: $k=148$; French: $k=139$; age: $M=44.01$, $SD=14.44$, range: 19–77; gender: 61% female; 38% male; 1% other). Although all participants in the final sample were L1 English speakers, 10% of the respondents reported having been exposed to one or more other languages at home in early childhood. Participants' demographic and language-related characteristics by group are summarized in Table 1.

Table 1. Participant characteristics

Characteristic	French ($k=139$)		Spanish ($k=148$)	
	k	%	k	%
Age				
Mean	44.24		43.78	
SD	14.53		14.40	
Range	20–77		19–74	
Gender				
Male	49	35	59	40
Female	90	65	85	57
Other			4	3
Native language(s)				
English	121	87	137	93
English + Other(s)	18	13	11	7
Other languages spoken (excluding French/Spanish)				
No	90	65	110	74
Yes	49	35	39	26
Reasons for learning French/Spanish ^a				
For travel	84	24	70	18
For school	4	1	7	2
For job-related purposes	18	5	40	10
For fun/leisure	121	34	112	28
For memory/brain acuteness	79	22	77	19
For social purposes	31	9	73	18
Other	16	5	21	5

Table 1. (continued)

Characteristic	French (<i>k</i> = 139)		Spanish (<i>k</i> = 148)	
	<i>k</i>	%	<i>k</i>	%
Other languages studied (excluding French/Spanish)				
No	31	22	34	23
Yes	108	78	114	77
Education ^b				
Some high school			1	1
High school	12	8	10	7
Associate's degree	4	3	10	7
Bachelor's degree	55	38	58	39
Master's degree	40	28	55	37
Ph.D.	18	13	6	4
Trade School	4	3	2	1
Other	11	8	7	5
Ethnicity				
Asian	7	5	4	3
African American	4	3	7	5
Caucasian	117	84	126	85
Latino or Hispanic	7	5	7	5
Other	4	3	4	3
Self-rated level of French/Spanish when started using the app ^c				
Mean	2.38		2.52	
SD	1.46		1.53	
Self-rated level of French/Spanish at pretest ^c				
Mean	4.18		4.29	
SD	1.31		1.56	
Self-rated level of French/Spanish at posttest ^d				
Mean	4.43		4.44	
SD	1.37		1.43	
Skills learned through Duolingo the most ^e				
Vocabulary	112	19	120	20
Grammar	87	15	102	17
Pronunciation	71	12	56	9
Listening	96	16	87	14

Table 1. (continued)

Characteristic	French (<i>k</i> =139)		Spanish (<i>k</i> =148)	
	<i>k</i>	%	<i>k</i>	%
Speaking	55	9	56	9
Reading	104	18	114	19
Writing	69	12	80	13
Using Duolingo resources other than regular lessons ^f				
Stories	110	49	111	46
Podcasts	36	16	34	14
Tips	64	29	73	30
Nothing	14	6	25	10
Experience learning French/Spanish before using Duolingo				
No	42	30	32	22
Yes	97	70	116	78
Ways of learning French/Spanish before using Duolingo ^g				
Being around native speakers	17	9	40	18
High school classes	65	36	79	36
Language apps	41	23	46	21
Internet-based materials (e.g., podcasts, YouTube)	9	5	11	5
Textbooks and other materials in print	3	2	4	2
Conversational language classes	25	14	22	10
Other	21	12	20	9
Taking French/Spanish classes in addition to Duolingo				
No	134	96	140	95
Yes	5	4	8	5
Using other programs/apps in addition to Duolingo				
No	116	83	130	88
Yes	23	17	18	12

Notes.

- a. French: *k* = 353, Spanish: *k* = 400.
b. French: *k* = 144, Spanish: *k* = 149.
c.d o = Absolute beginner; 10 = Native speaker.
e. French: *k* = 594, Spanish: *k* = 615.
f. French: *k* = 224, Spanish: *k* = 243.
g. French: *k* = 181, Spanish: *k* = 222.

3.2 Instruments and materials

Three different types of data and corresponding data sources were used in the study.

1. *Exposure/behavioral data*

The first data source shed light on learners' exposure to the target language and related behavior (in-app engagement). Of particular interest was (a) the duration of app usage measured as total minutes per participant across the 6-month period of study (i.e., "Minutes"), (b) the number of times the learner opened the app in a given week (i.e., "Logins") and the number of days the learner completed at least one lesson (i.e., "Sessions") – two frequency measures, and (c) the following content-related/curriculum-oriented intensity variables: "Lessons" (i.e., the number of lessons completed), "Level reviews" (i.e., the final lesson for a given Level/Skill combination), "Skill practice" (i.e., when a learner goes back to review skills that they have already "gilded"), "Stories" (i.e., the number of stories completed), and "Tests" (i.e., the number of tests completed). All of these indicators were used in their raw forms as predictors of learner gains.

2. *Self-report data*

To understand learner demographics as well as participants' language learning history, an instrument that largely mirrored Jiang, Rollinson, Plonsky, et al.'s (2021) background questionnaire was used. Additionally, we collected data using scales for measuring two individual difference variables: L2 grit (adapted from Teimouri et al., 2022) and L2 motivated learning behavior (adapted from Papi et al., 2019). These variables, individually and in tandem, allowed the study to examine these two individual differences as additional predictors of both in-app engagement and gains in learning.

Teimouri et al. (2022) validated their instrument with a sample of 191 learners of English in a foreign language context (Iran) and reported Cronbach's α reliability of .80 for the full L2 grit scale, .86 for the perseverance of effort subscale (PE, five positively keyed items), and .66 for the consistency of interest subscale (CI, four negatively keyed items). Papi et al.'s (2019) L2 motivated learning behavior questionnaire (five positively keyed items) was first used with a sample of 257 learners of English in a second language context (the US) and had internal consistency-reliability of .86 as measured by Cronbach's α .

Both scales were employed after implementing minor adjustments. Specifically, the word *English* in the original scales was replaced with French or Spanish in the present study in order to tailor the item wording to participants' target languages. Additionally, for the sake of consistency, Papi et al.'s (2019) original 5-point

Likert-type scale (endpoints: 1=*never true of me*; 5=*always true of me*) was replaced with Teimouri et al.'s (2022) 5-point fully verbal and numerical Likert-type scale (endpoints: 1=*not like me at all*; 5=*very much like me*). An example item for L2 grit: "I am a diligent French/Spanish learner"; an example item for L2 motivated learning behavior: "I work hard at studying French/Spanish."

3. *L2 proficiency*

Two different types of language tests were used to measure participants' L2 proficiency: A C-test (Spanish: Riggs & Maimone, 2018; French: Counsell, 2018) and an elicited imitation test (EIT; Spanish: Solon et al., 2019; French: Gaillard & Tremblay, 2016). These instruments were chosen based on a number of considerations. First, Spanish and French C-tests and EITs have undergone rigorous development and possess strong validity arguments. To illustrate, Riggs and Maimone (2018) reported a high correlation between Spanish C-test scores and (a) self-assessed proficiency ($r=.81, p<.001$) as well as (b) class level ($r=.73, p<.001$). Counsell (2018) also reported sizeable and positive correlations between French C-test scores and (a) self-assessed proficiency ($r_s=.58-.67$ for reading, writing, listening, and speaking, respectively, with an overall $r=.63, p<.01$) and (b) program level of study at the university ($r=.85, p<.01$). In the same vein, Gaillard and Tremblay (2016) found that the strongest predictors of French EIT ratings in their study were (a) C-test scores ($R^2=.79$) and (b) class level ($R^2=.69$). Solon et al. (2019), in turn, suggested that "the modified, 36-item EIT is, in fact, better able to discriminate among learners at higher levels of proficiency than is the 30-item EIT" (p.14) and reported Cronbach's α reliability ranging from .78 to .97 for the 30-item EIT and from .84 to .97 for the 36-item EIT for L2 learners at different proficiency levels. Another benefit of using C-tests and EITs is that they are considered to be good measures of explicit and implicit L2 knowledge, respectively (e.g., Ellis, 2005; Heo, 2016).³

Moreover, the validity of these two groups of tests is supported not only by primary studies but also by two recent meta-analyses. Synthesizing results across 239 studies, McKay (2019) found an almost perfect correlation between C-tests and tests of general language proficiency ($r=.94$). The evidence for EITs is likewise very strong. Kostromitina and Plonsky's (2022) meta-analysis observed an attenuation corrected correlation of $r=.81$ between EITs and other largely standardized tests of L2 proficiency.

Second are a set of practical considerations. These proficiency measures are highly efficient and can be completed independently and online in approximately 20–30 minutes. Upon completion, these tests can then be scored quickly and

3. We thank another anonymous reviewer for pointing this out.

accurately. The instruments are also freely available and do not carry any proprietary restrictions.

Finally, both C-tests and EITs have been developed and are available in a range of languages (Arabic, Chinese, German, Japanese, Russian). Therefore, the present study could be replicated in other L2s without changing this critical design feature (i.e., the dependent measure). The instruments used to collect self-report and L2 proficiency data are available in Appendix A.

3.3 Procedure

Following IRB approval, the self-report and L2 proficiency measures were pilot-tested, and the data were collected using an online survey platform Gorilla (<https://gorilla.sc>). Eligible app users (see above) were invited to be part of the study starting on August 3, 2021; the first round of data collection lasted until November 5, 2021. Those who expressed interest were asked to begin the study by completing an online survey that included a consent form followed by (a) instruments for L2 grit and motivation (items for each scale were randomized to control for order effect), (b) a language background questionnaire, (c) an EIT, and (d) a C-test for their chosen language (Spanish or French). This battery of instruments was completed remotely, without a proctor, in about an hour. Upon completion, participants were reminded of the minimum app engagement required for participation (i.e., at least 2 logins to the app per week for the following 26 weeks).

Six months later (i.e., in February through May 2022), each participant who had met the eligibility requirement was contacted again and invited to retake the two individual difference scales as well as the two proficiency tests. Participants who met the selection criteria and the eligibility requirements received a \$ 100 Amazon gift card. The selection criteria included: (a) being a Duolingo user studying either Spanish or French and (b) being a native speaker of English residing in the US. The requirements included: (a) completion of a survey and two language tests (at the time of the pretest and posttest 6 months later) and (b) a minimum of 52 logins on the Duolingo app (2 per week \times 26 weeks) to ensure minimally sufficient engagement with the target language and Duolingo content.

After all data were collected and de-identified, the C-tests were scored automatically, whereas the EITs were scored by four trained raters, all highly proficient in the target language (two raters per language: the lead rater scored all of the EIT items, whereas the second rater scored approximately 10% of the sample's EITs). Following rater training and norming sessions, which lasted approximately two hours, the raters for each language (French vs. Spanish) got calibrated themselves and proceeded to independently score the EITs. To avoid potential rater

bias, raters were kept unaware of which audio files had come from the pretest and which were from the posttest.

3.4 Data analysis

To calculate interrater reliability for the EIT scores, intraclass correlation coefficients (ICC, two-way mixed, consistency) were computed: (a) French: average α for 28 items = .96; average by test type (14 items each): pretest = .95, posttest = .96; (b) Spanish: average α for 30 items = .98; average by test type (15 items each): pretest = .99, posttest = .97.

During the data clean-up, 30 cases (10% out of a total of 287; 17 Spanish, 13 French) were excluded listwise due to issues with recordings or participants' misinterpretation of the task (several produced English translations rather than French/Spanish imitations). There were no missing data on other variables.

To compare proficiency scores across languages and proficiency test modes (written vs. oral), they were first converted to decimals separately by language. Spanish EIT: 36 items, max possible score = 144 (4 per item). French EIT: 50 items, max possible score = 200 (4 per item). C-test (both languages): 125 items, max possible score = 125 (1 per item). Next, the assumptions for each statistical analysis were checked and met (see Appendix B).

4. Results

4.1 Preliminary analyses: Scale data

The inspection of the scale data revealed two items with low corrected item-total correlations (ITCs < .40) on the L2 Grit Consistency of Interest subscale: CI7R and CI8R, which considerably affected reliability of the scale and were, therefore, removed from further analyses. The rest of the corrected ITCs for all constructs and subconstructs were > .40 on both the pretest and the posttest. The stability of constructs over time was assessed by test-retest reliability: $r(\text{L2 grit}) = .68$; $r(\text{L2 perseverance of effort}) = .69$; $r(\text{L2 consistency of interest}) = .58$; $r(\text{L2 motivation}) = .61$, $p < .001$. As demonstrated in Table 2, internal-consistency reliability of the scales was also acceptable. Additionally, descriptive statistics indicated that the participants had the highest mean score on L2 consistency of interest and the lowest mean score on L2 motivation on both the pretest and the posttest.

Table 2. Reliability analyses for scales ($N=287$)

Variable	<i>k</i>	<i>M</i> <i>SD</i> α 95% CI				<i>M</i> <i>SD</i> α 95% CI			
		Pretest				Posttest			
L2 grit	7	3.96	0.62	.84	[.81, .87]	3.79	0.66	.87	[.84, .89]
L2 perseverance of effort	5	3.71	0.73	.86	[.83, .88]	3.57	0.72	.86	[.83, .88]
L2 consistency of interest	2	4.58	0.63	.78 ^a	[.72, .82]	4.35	0.77	.81 ^b	[.76, .85]
L2 motivation	5	3.55	0.77	.88	[.86, .90]	3.30	0.81	.89	[.87, .91]

Notes.

a. *k* = number of items; *M* = mean; *SD* = standard deviation; 95% CI = 95% confidence intervals of coefficient alphas; L2 = second language. Spearman's rho = .59

b. *k* = number of items; *M* = mean; *SD* = standard deviation; 95% CI = 95% confidence intervals of coefficient alphas; L2 = second language. Spearman's rho = .65

4.2 Preliminary analyses: Proficiency data

As shown in Table 3, the participants' mean scores on the EIT (i.e., oral proficiency test) were higher in Spanish than in French, which was observed during both the pretest and the posttest. However, Spanish EIT scores were more spread out, as demonstrated by higher standard deviations. The participants' average C-test scores (i.e., written proficiency test) were overall slightly higher than the EIT scores, except for Spanish pre-test mean scores, which were virtually the same in both modes. Nonetheless, the two groups' proficiency in the written mode appeared to be at about the same level on both the pretest and the posttest (see Figure 1). The results of dependent-samples *t*-tests showed that learner proficiency gains from the pretest to the posttest were significant, with small effect sizes adjusted for the within-sample correlation (Plonsky & Oswald, 2014): (a) EIT gains: $t(256)=13.38$, $p<.001$, Cohen's $d=.29$, 95% CI [.24, .33]; (b) C-test gains: $t(286)=11.00$, $p<.001$, Cohen's $d=.36$, 95% CI [.29, .43].

Table 3. Descriptive statistics for proficiency test scores ($N=287$)

Proficiency Tests	<i>M</i>	<i>SD</i>	Min	Max
EIT pretest French	.12	.08	.005	.47
EIT posttest French	.17	.11	.03	.57
EIT pretest Spanish	.27	.15	.02	.77
EIT posttest Spanish	.30	.16	.01	.84
C-test pretest French	.28	.13	.02	.68
C-test posttest French	.32	.14	.00	.78
C-test pretest Spanish	.27	.12	.00	.68
C-test posttest Spanish	.32	.12	.07	.75

Notes. *M* = mean; *SD* = standard deviation. EIT (elicited imitation test) = oral proficiency; C-test = written proficiency.

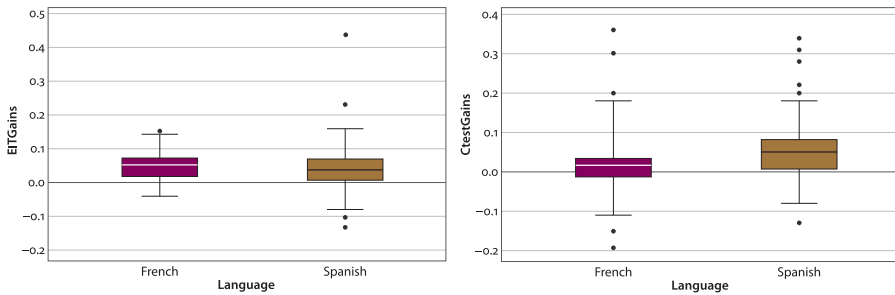


Figure 1. Proficiency test gains in the written vs. oral mode and by language

4.3 RQ1: To what extent do learner gains differ when tested in the written vs. oral mode?

The results of dependent-samples *t*-tests for RQ1 demonstrated that there were no statistically significant differences in learner proficiency gains in the written vs. oral mode (see also Figure 1), and the effect sizes adjusted for the within-sample correlation were small (see Plonsky & Oswald, 2014): (a) French: $t(125) = -.71$, $p = .48$, Cohen’s $d = -.07$, 95% CI $[-.27, .13]$; (b) Spanish: $t(130) = 1.23$, $p = .22$, Cohen’s $d = .13$, 95% CI $[-.08, .34]$.

As shown in Figure 1, the results of the independent-samples *t*-tests for RQ1 showed that the EIT gains (oral mode) did not statistically differ by language: $t(204.64) = -.85$, $p = .397$, Cohen’s $d = -.11$, 95% CI $[-.35, .14]$. The same was true for the C-test gains (written mode) by language: $t(285) = 1.52$, $p = .13$, Cohen’s $d = .18$,

95% CI [-.05, .41]. Of note, similar results were observed when the two tests were re-run without outliers on the dependent variable.

4.4 RQ2: To what extent are frequency, duration, and intensity of Duolingo app usage associated with gains in L2 Spanish and French?

A summary of descriptive statistics for the Duolingo app usage data is demonstrated in Table 4. It shows meaningful differences among the frequency, duration, and intensity variables. (A slight overlap between the two frequency variables was addressed when conducting a follow-up multiple regression analysis, see Table 6.)

Table 4. Descriptive statistics for in-app usage data

Type	App usage	<i>M</i>	<i>SD</i>	Min	Max
Frequency	Logins	164.1	13.57	105	180
	Sessions	154.7	18.82	50	180
Duration	Minutes	3070	2812	252.7	22,753.8
Intensity: Content and Curriculum	Lessons	527.0	1395	9.0	19,701
	Level reviews	45.06	46.81	0	309
	Skill practice	93.22	157.4	0	1434
	Stories	170.9	809.3	0	13,346
	Tests	19.99	85.74	0	854

As demonstrated in Table 5, learners' self-determined exposure and behavior (i.e., frequency, duration, and intensity of Duolingo app usage) were weakly to moderately correlated with their language proficiency gains made over a six-month period. More specifically, the exposure-related variables most strongly associated with gains measured in the oral model (i.e., via EIT) were the number of lessons, level reviews, and logins. Raw duration measured in total minutes of exposure during the 6-month period of study exhibited almost no association with gains in the oral mode. As with EIT gains, in-app exposure/behaviors associated with gains in proficiency in the written mode included the number of lessons completed and total number of level reviews (although the latter association was neither particularly strong nor statistically significant, as shown by a confidence interval that crossed zero). Unlike EIT gains, however, minutes of exposure was found to be associated with C-test gains.

While addressing RQ2, it occurred to us that some of the observed correlations may have been artificially attenuated due to a restricted range of observed

values. In particular, requiring at least two logins per week, though necessary to ensure regular exposure to the target language, seems to have yielded an unusual level of homogeneity in login data across the sample: The mean number of logins was 165 (6.35 logins per week) with a standard deviation of only 11.30. The relationships between logins and gains on the EIT and C-test were then re-examined using Thorndike's formula for correction for range restriction, resulting in substantially larger correlations in both cases (i.e., .46 and .11, respectively).

Table 5. Pearson correlations between Duolingo app usage and proficiency gains ($N=233$)

Type	App usage	EIT gains	C-test gains
Frequency	Logins ^a	$r=.14$, BCa 95% CI [.03, .26]*	$r=.03$, BCa 95% CI [-.09, .17]
	Sessions	$r=.12$, BCa 95% CI [.00, .24]	$r=.06$, BCa 95% CI [-.07, .19]
Duration	Minutes	$r=.01$, BCa 95% CI [-.16, .18]	$r=.20$, BCa 95% CI [.05, .35]
Intensity: Content and Curriculum	Lessons	$r=.21$, BCa 95% CI [.07, .35]	$r=.26$, BCa 95% CI [.13, .40]
	Level reviews	$r=.21$, BCa 95% CI [.07, .34]	$r=.11$, BCa 95% CI [-.04, .27]
	Skill practice	$r=-.06$, BCa 95% CI [-.19, .07]	$r=.08$, BCa 95% CI [-.04, .22]
	Stories	$r=-.02$, BCa 95% CI [-.16, .14]	$r=.07$, BCa 95% CI [-.04, .18]
	Tests	$r=.09$, BCa 95% CI [-.02, .25]	$r=.09$, BCa 95% CI [-.06, .22]

Notes.

* Bias-corrected and accelerated 95% confidence intervals for correlation coefficients.

a. When adjusted for range restriction, the correlations for logins x EIT ($r=.14$) and logins x C-test gains ($r=.03$) increase to .46 and .11, respectively.

The results of the first multiple regression analysis suggested that when the three variables most strongly associated with EIT gains were entered into the model as predictors along with the target language, which was added as a covariate, the model explained 4–6% of the variance in EIT gains and was statistically significant: $F(4,238)=3.56$, $p=.008$, $R^2=.06$, adjusted $R^2=.04$. ‘Level Reviews’ emerged as the only meaningful positive predictor (see Table 6).⁴

The results of the second multiple regression analysis demonstrated that when the two variables most strongly correlated with C-test gains were entered in the model as predictors along with the target language, which was added as a covariate, the model explained 5–6% of the variance in C-test gains and was statistically significant: $F(3,257)=5.21$, $p=.002$, $R^2=.06$, adjusted $R^2=.05$. ‘Lessons’ emerged as the only meaningful positive predictor (see Table 7).

4. Of note, the Sessions variable was excluded from the model due to a large correlation with the Logins variable.

Table 6. Regression analysis summary for variables predicting EIT gains

Variable	<i>B</i>	SE	β	<i>t</i>	<i>p</i>
Target language	-.008	.006	-.095	-1.49	.14
Logins	.000	.000	.059	.887	.38
Lessons	8.820e-6	.000	.051	6.58	.51
Level reviews	.000	.000	.158	2.052	.04

Notes. *N* = 243. Overall model: $R^2 = .06$, adjusted $R^2 = .04$.

Table 7. Regression analysis summary for variables predicting C-test gains

Variable	<i>B</i>	SE	β	<i>t</i>	<i>p</i>
Target language	.012	.007	.101	1.651	.10
Minutes	7.784e-7	.000	.024	.255	.80
Lessons	5.152e-5	.000	.204	2.127	.03

Notes. *N* = 261. Overall model: $R^2 = .06$, adjusted $R^2 = .05$.

4.5 RQ3: To what extent are L2 grit and motivation associated with the frequency, duration, and intensity with which learners use Duolingo?

As shown in Table 8, the individual difference variables of L2 grit and motivation measured at the pretest were weakly to moderately correlated with learners’ frequency, duration, and intensity of the Duolingo app usage. Although the magnitude of effect sizes was small, meaningful positive relationships were observed between the two frequency variables (i.e., “Logins” and “Sessions”) and L2 grit perseverance of effort as well as L2 motivation. Additionally, duration (i.e., “Minutes”) and content-related intensity (i.e., “Lessons” and “Stories”) were positively correlated with the two subcomponents of L2 grit as well as L2 motivation.

Table 8. Pearson correlations between Duolingo app usage and individual differences (*N* = 260)

Type	App usage	L2 Perseverance of effort	L2 Consistency of interest	L2 Motivation
Frequency	Logins	$r = .16$ [.06, .27]*	$r = .08$ [-.05, .20]	$r = .20$ [.08, .31]
	Sessions	$r = .15$ [.04, .26]	$r = .12$ [-.02, .23]	$r = .20$ [.07, .31]
Duration	Minutes	$r = .18$ [.05, .30]	$r = .13$ [.02, .22]	$r = .24$ [.12, .35]

Table 8. (continued)

Type	App usage	L2 Perseverance of effort	L2 Consistency of interest	L2 Motivation
Intensity: Content and curriculum	Lessons	$r = .17$ [.02, .31]	$r = .13$ [.03, .22]	$r = .22$ [.09, .33]
	Level reviews	$r = -.002$ [-.11, .11]	$r = .01$ [-.11, .11]	$r = .07$ [-.05, .18]
	Skill practice	$r = .003$ [-.12, .13]	$r = -.01$ [-.13, .11]	$r = .03$ [-.10, .15]
	Stories	$r = .16$ [.03, .28]	$r = .12$ [.02, .22]	$r = .17$ [.05, .27]
	Tests	$r = .09$ [-.04, .20]	$r = .03$ [-.09, .13]	$r = .06$ [-.05, .15]

Note.

* Bias-corrected and accelerated 95% confidence intervals for correlation coefficients.

4.6 RQ4: To what extent are L2 grit and motivation associated with gains in L2 Spanish and French?

Pearson correlations revealed the extent to which L2 grit and motivation measured at the pretest were associated with learner gains in written and oral proficiency. The observed relationships were positive and constituted generally small effect sizes (Plonsky & Oswald, 2014). As in response to some of the previously conducted analyses, we applied a correction to the correlations that we had reason to believe may have been attenuated due to range restriction. Specifically, the standard deviations observed for both of the L2 grit subconstructs were substantially smaller than in previous studies of foreign-language learners (e.g., Sudina & Plonsky, 2021b) and were adjusted accordingly yet conservatively, shown in the following results in parentheses following the corresponding uncorrected correlations: $r = .02$, BCa 95% CI [-.10, .16] between EIT gains and L2 perseverance of effort; $r = .12$ ($r_{\text{corrected}} = .22$), BCa 95% CI [-.01, .25] between EIT gains and L2 consistency of interest; $r = .03$, BCa 95% CI [-.10, .16] between EIT gains and L2 motivation; $r = .16$ ($r_{\text{corrected}} = .19$), BCa 95% CI [.04, .28] between C-test gains and L2 perseverance of effort (note that the confidence interval does not cross zero); $r = .09$ ($r_{\text{corrected}} = .17$), BCa 95% CI [-.04, .23] between C-test gains and L2 consistency of interest; $r = .17$, BCa 95% CI [.05, .30] between C-test gains and L2 motivation (note that the confidence interval does not cross zero).

To examine the extent to which individual differences of L2 grit and motivation predicted EIT/C-test gains, two standard multiple regressions with four predictors (i.e., L2 perseverance of effort, L2 consistency of interest, L2 motivation, and target language, which was added as a covariate) and EIT/C-test gains as outcome variables were performed. The two models explained 1–3% of the variance in gains scores and were not statistically significant: (a) $F(4, 244) = 1.76$,

$p = .14$, $R^2 = .03$, adjusted $R^2 = .01$ for EIT gains, with L2 consistency of interest as the only contributing predictor: $\beta = .14$, $B = .01$, 95% CI [.00, .03], $p = .05$; (b) $F(4, 268) = 1.83$, $p = .12$, $R^2 = .03$, adjusted $R^2 = .01$ for C-test gains, with no meaningful predictors.

5. Discussion

The current study sought to examine the predictive power of two sets of variables on L2 gains made in app-based language learning via Duolingo. Specifically, we were interested in better understanding L2 development as a function of both (a) learners' app-based exposure/behavior (e.g., instructional frequency, duration) as well as (b) learners' L2 grit and motivation. On a broad, theoretical level, these sets of variables represent the two main types of factors (learner-external and learner-internal) known to influence L2 learning (Gass et al., 2020). On a practical level, the results have the potential to inform the instructional design of Duolingo's curriculum and to provide implications for in-app experience that increase learner efficiency.

The study is unique in at least two respects. First, to our knowledge, this is the only study to consider distribution of practice effects in the context of mobile-based language learning. Moreover, we have done so by means of a natural experiment thereby greatly increasing the study's ecological validity. Second, although a growing body of evidence has begun to accumulate on the role of grit in L2 development (see Teimouri et al., 2021), no study to date has done so with mobile language learners. It is also the first study to employ a longitudinal design to examine the power of grit in predicting gains over time.

One challenge to these goals, which we want to be upfront about, were the relatively modest gains observed on both the written and oral proficiency tests (i.e., C-test and EIT) in both languages. The lack of target language gains that were observed over time (i.e., our main dependent variable) imposed a limitation on the study's findings because less gains necessarily means less for the predictor or independent variables to explain. These gains appear in conflict with previous findings on the effectiveness of Duolingo (e.g., Jiang et al., 2021). However, there are several alternate explanations. For example, unlike other standardized proficiency tests (e.g., ACTFL's Oral Proficiency Interview), the dependent measures in the present study were not developed with lower proficiency levels in mind and may have been too difficult, as noted to us by several participants. Another explanation for the modest gains observed may be a lack of effort on pre- and post-assessments on the part of learners. Finally, we need to account for the user autonomy and the amount of the course content covered by the participants after

six months of learning. A follow-up analysis of participants' maximum course tree depth (i.e., the furthest row of each section of the courses the participants were at) within seven days of completing the posttest suggested that our participants did not cover enough new content over six months of learning on Duolingo. The results revealed that the majority of our learners remained at the beginning level at the posttest based on the amount of material they covered. This finding, along with considerable attrition rates in the present study, echoes Loewen et al.'s (2019) observation that "learners may not persist long enough to make considerable gains in their L2 knowledge, especially without any obligation or encouragement from peers and teachers commonly found in classroom environments" (p.308). In light of this finding, one suggestion for efficient use of Duolingo is moving forward and learning new content despite the challenges along the way. To increase motivation and accountability, Duolingo users might consider adding and following friends and peers who are also using the app.

As stated above, one of our main interests in this study was to examine language gains made in relation to learners' in-app exposure (frequency, duration, intensity) and associated behaviors (e.g., content-related choices). Somewhat surprisingly, total minutes of exposure during the period of study only exhibited a correlation with gains when measured with the C-test. Rather, the more consistent and dependable (i.e., across modes and dependent measures) correlates of gains were the more frequency- and curriculum-oriented measures such as the number of lessons, reviews, and logins.

This finding carries several important ramifications and interpretations. First, these results generally confirm the lack of a relationship between hours of exposure on Duolingo and language gains observed in Jiang et al. (2021). Such a finding might be perceived as counter-intuitive or even disappointing in that greater time spent by learners does not necessarily yield greater gains. However, we view these findings more optimistically in that they indicate that gains can be made even with shorter, more frequent and purposeful in-app engagement. Moreover, the lack of a relationship between gains and time spent using the app is even more noteworthy given the relative homogeneity of logins across the participants. In other words, although the participants were free to log in as frequently as they liked as long as they did so at least twice per week during the 26-week period of study, most logged in on a daily or almost daily basis (>6 logins/week), thus providing a kind of built-in control for the role of frequency of exposure. This finding also aligns well with Kim and Webb's recent (2022) meta-analysis showing a marked advantage for more frequent exposure to the target language as opposed to longer (in minutes, hours) periods of exposure; in other words, the shorter spacing gaps showed greater frequency effects. From a skill-acquisition perspective, this finding could be explained in terms of the intensity of exposure favoring

proceduralization processes (see Serrano, 2011). Another study that demonstrated a smaller frequency effect in spaced learning conditions compared to massed conditions was Uchihara et al.'s (2019) meta-analysis of repetition effects in incidental vocabulary learning. The major implication here for learners is fairly straightforward: Log in to the app frequently with the goal of completing entire lessons and reviews, even if those sessions do not last for long periods of time.

Moreover, the fact that the total number of level reviews was positively correlated with and emerged as a meaningful predictor of oral proficiency gains (RQ2) deserves special attention. Based on the cognitive psychology definition of spacing (Cepeda et al., 2008, see the literature review for more), a true spacing study must involve the repetition of the same stimuli from the first study session to the next. Level review appears to be where the participants return to review the same content. Our findings indicate that engagement in level reviews is positively related to and serves as a predictor of learning gains in the oral mode, which is in line with the literature on distribution of practice effects (Nakata, 2015).

Research questions 3 and 4 both involved the two learner individual differences variables of L2 grit and motivation. RQ3 was concerned with the association between these two variables and the learners' in-app behaviors and exposure. It is natural to expect that learners who possess higher levels of (L2) grit and/or motivation would engage more often and/or more thoroughly with the Duolingo curriculum. This supposition was indeed the case at least for some of the individual differences and for some of the learner behaviors. Not surprisingly, motivation exhibited some of the strongest associations with in-app exposure as well as written proficiency gains; the latter is congruous with other app-based language learning studies (e.g., Loewen et al., 2020 found that learner motivation was a significant predictor of oral proficiency scores for learners who had studied Spanish using Babbel for 12 weeks). However, we also observed meaningful and statistically significant correlations between L2 grit and several in-app behaviors and proficiency gains. In fact, L2 grit consistency of interest (and not motivation as in Loewen et al., 2020) was the only contributing predictor of oral proficiency gains.

The findings for RQ3 are noteworthy on multiple levels. First of all, we have to understand that, on a theoretical level, grit and motivation do not in and of themselves induce greater learning. Rather, as this study shows, these qualities are associated with the types of activities that lead to learning such as seeking out target language interlocutors, spending time studying or reading in the target language, or – most pertinent to the present study – engaging with the target language instruction via a language learning app such as Duolingo. Thus, the findings of the present study provide one of the first pieces of evidence of the predictive validity of the L2 grit scale of language learning behaviors. From an educational standpoint, these results demonstrate that greater motivation and grit may

lead to higher frequency of the types of activities shown in response to RQ2 to be associated with language gains (see Figure 2). Duolingo may be interested, therefore, in seeking to foster learner motivation and grit as a means to enhance linguistic development if only indirectly.

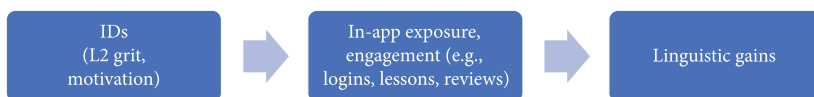


Figure 2. Hypothesized model of the effect of individual differences on app-based learning, mediated by in-app exposure and engagement

RQ4 addressed the same relationship modeled in Figure 2 but without the mediating effect of in-app exposure. The findings for this relationship were modest but provide additional evidence of the predictive validity of L2 grit in the context of app-based language learning.

6. Conclusion

The findings of this study carry relevance and potential benefits on multiple levels. First, this study allowed us to gain a better understanding of the role of technology in instructed second-language acquisition. This is critical as technological advances have the potential to make language learning not only “a lifelong (spanning one’s lifetime) but also a lifewide (not confined to a particular location, such as a school) activity” (Reinders & Stockwell, 2017, p.372). Second, the present investigation contributed to the growing line of evidence of Duolingo’s effectiveness by assessing L2 learners’ proficiency in both written and oral modes using high validity and high practicality measures. The study also shed further light on our understanding of the individual and combined effects of frequency, duration, and intensity of instruction on L2 development and, critically, on the learner-internal factors that lead to choices to engage with the app. Finally, on a practical level, the results of the present study may also inform Duolingo lesson design and recommendations provided to learners with respect to the frequency, duration, and intensity of app usage.















Acknowledgements

This project was funded by a Duolingo Efficacy Study grant. We are very grateful to the Learning Science team at Duolingo for seeing the value in this study and for all their support and assistance. In particular we would like to thank Xiangying Jiang, Erin Gustafson, and Joseph

Rollinson for their patience, generosity, and help digging up the learner-usage data we needed time and time again. We are also very grateful to the language learners who contributed their time and energy (and data) to this study. In addition, our sincere thanks go to Kevin Hirschi, Masha Kostromitina, Ben Brown, and Andrew Dennis, for their tireless assistance with scoring, coding, piloting, and Gorilla-wrangling (the data collection platform, not the primate). Thanks to Kate Yaw for help recording our test instructions. Last, our gratitude goes out to the C-test and elicited imitation test authors who very kindly provided us with the materials needed to employ their instruments in this study: Stéphanie Gaillard, Annie Tremblay, Corinne Counsell, Daniel Riggs, Luciane L. Maimone, Megan Solon, Hae In Park, Carly Henderson, and Marzieh Dehghan-Chaleshtori.

References

- doi Carpenter, S. K. (2020). Distributed practice/spacing effect. In Li-fang Zhang (Ed.), *Oxford Research Encyclopedia of Education* (pp. 1–20). Oxford University Press.
- doi Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridge of optimal retention. *Psychological Science*, *19*(11), 1095–1102.
- Council of Europe. (2001). *Common European Framework of Reference for languages: Learning, teaching, assessment*. Retrieved from <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. 203–230). Peter Lang.
- doi Craig, P., Katikireddi, S. V., Leyland, A., & Popham, F. (2017). Natural experiments: An overview of methods, approaches, and contributions to public health intervention research. *Annual Review of Public Health*, *38*, 39–56.
- doi Delaney, P. F., Verkoeijen, P. P. J. L., & Spigel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 53, pp. 63–147). Academic Press.
- doi Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long term goals. *Journal of Personality and Social Psychology*, *92*, 1087–1101.
- doi Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A Psychometric Study. *Studies in Second Language Acquisition*, *27*(2), 141–172.
- doi Gaillard, S., & Tremblay, A. (2016). Linguistic proficiency assessment in second Language acquisition research: The elicited imitation task. *Language Learning*, *66*, 419–447.
- doi 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*, 71–97.
- doi Gass, S. M., Behney, J., & Plonsky, L. (2020). *Second language acquisition: An introductory course* (5th ed.). Routledge.
- Heo, Y. (2016). Heritage and L2 learners' acquisition of Korean in terms of implicit and explicit knowledge. PhD dissertation, Michigan State University, East Lansing.

- Jiang, X., Chen, H., Portnoff, L., Gustafson, E., Rollinson, J., Plonsky, L., & Pajak, B. (2021). Seven units of Duolingo courses comparable to 5 university semesters in reading and listening. *Duolingo Research Report DRR-21-03*.
- Jiang, X., Rollinson, J., Chen, H., Reuveni, B., Gustafson, E., Plonsky, L., & Pajak, B. (2021). How well does Duolingo teach speaking skills? *Duolingo Research Report DRR-21-02*.
- Jiang, X., Rollinson, J., Plonsky, L., & Pajak, B. (2020). Duolingo efficacy study: Beginning-level courses equivalent to four university semesters. *Duolingo Research Report DRR-20-04*.
-  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, 974–1002.
-  Kasprowicz, R. E., Marsden, E., & Sephton, N. (2019). Investigating distribution of practice effects for the learning of foreign language verb morphology in the young learner. *Modern Language Journal*, 103, 580–606.
-  Kim, S. K., & Webb, S. (2022). The effects of spaced practice on second language learning: A meta-analysis. *Language Learning*, 72, 269–319.
-  Kostromitina, M., & Plonsky, L. (2022). Elicited imitation tasks as a measure of L2 proficiency: A meta-analysis. *Studies in Second Language Acquisition*, 44, 886–911.
-  Li, M., & DeKeyser, R. (2019). Distribution of practice effects in the acquisition and retention of L2 Mandarin tonal word production. *Modern Language Journal*, 103, 607–628.
-  Loewen, S. (2020). *Introduction to instructed second language acquisition* (2nd ed.). Routledge.
-  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.
-  Loewen, S., Isbell, D., & 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.
- McKay, T. (2019). More on the validity and reliability of C-test scores: A meta-analysis of C-test studies (Unpublished doctoral dissertation). Georgetown University, Washington, DC.
-  Nakata, T. (2015). Effects of expanding and equal spacing on second language vocabulary learning: Does gradually increasing spacing increase vocabulary learning?. *Studies in Second Language Acquisition*, 37, 677–711.
-  Nakata, T., & Suzuki, Y. (2019). Effects of massing and spacing on the learning of semantically related and unrelated words. *Studies in Second Language Acquisition*, 41, 287–311.
-  Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, 50, 417–528.
-  Papi, M., Bondarenko, A., Mansouri, S., Feng, L., & Jiang, C. (2019). Rethinking L2 motivation research: The 2 × 2 model of L2 self-guides. *Studies in Second Language Acquisition*, 41, 337–361.
-  Park, D., Yu, A., Baelen, R. N., Tsukayama, E., & Duckworth, A. L. (2018). Fostering grit: Perceived school goal-structure predicts growth in grit and grades. *Contemporary Educational Psychology*, 55, 120–128.
-  Pawlak, M., & Kruk, M. (2022). *Individual differences in computer assisted language learning research*. Routledge.

- doi Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, 64, 878–912.
- Plonsky, L., & Ziegler, N. (2016). The CALL-SLA interface: Insights from a second-order synthesis. *Language Learning & Technology*, 20, 17–37.
- doi Reinders, H., & Stockwell, G. (2017). Computer-assisted SLA. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 108–125). Routledge.
- Riggs, D., & Maimone, L. L. (2018). A computerized-administered C-test in Spanish. In J. M. Norris (Ed.), *Developing C-tests for estimating proficiency in foreign language research* (pp. 265–294). Peter Lang.
- doi Rohrer, D. (2015). Student instruction should be distributed over long time periods. *Educational Psychology Review*, 27(4), 635–643.
- doi Rogers, J. (2017). The spacing effect and its relevance to second language acquisition. *Applied Linguistics*, 38(6), 906–911.
- doi Rogers, J. (2023). Spacing effects in task repetition research. *Language Learning*, 73(2), 445–474.
- doi Rogers, J., & Cheung, A. (2020). Input spacing and the learning of L2 vocabulary in a classroom context. *Language Teaching Research*, 24(5), 616–641.
- doi Rogers, J., & Cheung, A. (2021). Does it matter when you review? Input spacing, ecological validity, and the learning of L2 vocabulary. *Studies in Second Language Acquisition*, 43, 1138–1156.
- doi Saito, K., & Plonsky, L. (2019). Measuring the effects of second language pronunciation teaching: A proposed framework and meta-analysis. *Language Learning*, 69, 652–708.
- doi Schmidt, F. T., Nagy, G., Fleckenstein, J., Möller, J., & Retelsdorf, J. (2018). Same same, but different? Relations between facets of conscientiousness and grit. *European Journal of Personality*, 32, 705–720.
- doi Serrano, R. (2011). The time factor in EFL classroom practice. *Language Learning*, 61(1), 117–145.
- doi Serrano, R. (2022). A state-of-the-art review of distribution-of-practice effects on L2 learning. *Studies in Second Language Learning and Teaching*, 12(3), 355–379.
- doi Serrano, R., & Huang, H. Y. (2018). Learning vocabulary through assisted repeated reading: How much time should there be between repetitions of the same text? *TESOL Quarterly*, 52(4), 971–994.
- doi Serrano, R., & Huang, H. Y. (2023). Time distribution and intentional vocabulary learning through repeated reading: A partial replication and extension. *Language Awareness*, 32(1), 1–18.
- doi Solon, M., Park, H. I., Henderson, C., & Dehghan-Chaleshtori, M. (2019). Revisiting the Spanish elicited imitation task: A tool for assessing advanced language learners? *Studies in Second Language Acquisition*, 41, 1027–1053.
- doi Sudina, E., & Plonsky, L. (2021a). Academic perseverance in foreign language learning: An investigation of language-specific grit and its conceptual correlates. *Modern Language Journal*, 105, 829–857.
- doi Sudina, E., & Plonsky, L. (2021b). Language learning grit, achievement, and anxiety among L2 and L3 learners in Russia. *ITL – International Journal of Applied Linguistics*, 172, 161–198.

- doi Suzuki, Y. (2017). The optimal distribution of practice for the acquisition of L2 morphology: A conceptual replication and extension. *Language Learning*, 67(3), 512–545.
- doi Suzuki, Y. (2019). Individualization of practice distribution in second language grammar learning: The role of metalinguistic rule rehearsal ability and working memory capacity. *Journal of Second Language Studies*, 2(2), 169–196.
- doi Suzuki, Y., & DeKeyser, R. (2017). Effects of distributed practice on the proceduralization of morphology. *Language Teaching Research*, 21(2), 166–188.
- doi Teimouri, Y. (2017). L2 selves, emotions, and motivated behaviors. *Studies in Second Language Acquisition*, 39, 681–709.
- doi Teimouri, Y., Sudina, E., & Plonsky, L. (2021). On domain-specific conceptualization and measurement of grit in L2 learning. *Journal for the Psychology of Language Learning*, 3, 156–164.
- doi Teimouri, Y., Plonsky, L., & Tabandeh, F. (2022). L2 Grit: Passion and perseverance for second-language learning. *Language Teaching Research*, 26, 893–918.
- doi Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies. *Language Learning*, 69(3), 559–599.
- doi Yamagata, S., Nakata, T., & Rogers, J. (2023). Effects of distributed practice on the acquisition of verb-noun collocations. *Studies in Second Language Acquisition*, 45, 291–317.

Appendix A. The full instrument used in the study

(After signing the consent form). Thank you! Please set aside **60 minutes** of undisturbed time to participate in the study. You should have access to a **computer** with a **working microphone** and **speakers** and **reliable internet service**. Please, do **not** use a mobile phone.

Thank you! Please provide the information below:

Language learned on Duolingo: French/Spanish

Now think about your foreign language learning experience on Duolingo (in your case, French/Spanish) and respond to the following items by selecting the statements that best describe you. This is not a test and there are no right or wrong answers, so please be honest.

L2-GRIT SCALE

(adapted from Teimouri et al., 2022)

Perseverance of effort

1. I will not allow anything to stop me from making progress in learning French/Spanish*.
2. I am a diligent French/Spanish learner.
3. Now that I have decided to learn French/Spanish, nothing can prevent me from reaching this goal.
4. When it comes to French/Spanish, I am a hard-working learner.
5. I put much time and effort into improving my weaknesses in learning French/Spanish.

Consistency of interest

- 6R. I think I have lost my interest in learning French/Spanish.

- 7R. I have been obsessed with learning French/Spanish in the past but later lost interest.
- 8R. My interests in learning French/Spanish change from year to year.
- 9R. I am not as interested in learning French/Spanish as I used to be.

Note. ‘R’ indicates negatively keyed items that have been reversed.

*The original scale was developed for L2 English, which in the present study was replaced with *French/Spanish*.

L2 motivated learning behavior

(adapted from Papi et al., 2019)

- 1. I work hard at studying French/Spanish *.
- 2. I spend a lot of time studying French/Spanish.
- 3. I put a lot of effort in studying French/Spanish.
- 4. I constantly think about my French/Spanish learning activities.
- 5. Studying French/Spanish is very important to me these days.

Note. * The word *English* in the original scale was replaced with *French/Spanish* in the present study.

Response options and scoring

Not like me at all	Not much like me	Somewhat like me	Mostly like me	Very much like me
1	2	3	4	5

Note. The mean scores on each scale indicate the levels of L2 grit and L2 motivated learning behavior, respectively. The items were randomized.

Background questionnaire

(based on Jiang et al., 2021)

This form asks for background information about you. Although we ask for your name and email, we do so only because we want to associate your answers to this questionnaire with your other data. Your answers will be treated **confidentially**. Only the researchers will have access to the information you provide.

- 1. Name:
- 2. Email (the one that is associated with your **Duolingo** account):
- 3. Age (please put a number):
- 4. What language(s) was/were spoken in your home before you were 6 years old?
- 5. What other languages do you speak, if any?*
- 6. Why are you learning French/Spanish? (Check all that apply)

For travel	For school	For job-related purposes	For fun/leisure
For memory/brain acuteness	For social purposes	Other (please specify)	
- 7. What other languages have you studied?
- 8. What is your highest level of education?

Some high school	High school	Associate’s degree	Bachelor’s degree
Master’s degree	Ph.D.	Trade school	Other (please specify)
- 9. What gender do you identify as?

Male	Female	Other (please specify)
------	--------	------------------------
- 10. Please specify your ethnicity.

- Caucasian African American Latino or Hispanic
Asian Other (please specify)
11. How much French/Spanish do you think you knew on Duolingo when you started using the app?
0 1 2 3 4 5 6 7 8 9 10
Absolute beginner Native speaker
12. How much French/Spanish do you think you know now?
0 1 2 3 4 5 6 7 8 9 10
Absolute beginner Native speaker
13. In which area(s) do you think Duolingo helped you the most? (Check all that apply)
Vocabulary Grammar Pronunciation Listening
Speaking Reading Writing
14. How much time (in hours) per week did you use Duolingo to learn French/Spanish?
15. In addition to the Duolingo French/Spanish lessons, what other Duolingo resources did you use to learn the language? (Check all that apply)
Duolingo French/Spanish Stories
Duolingo French/Spanish Podcasts
Duolingo Tips in French/Spanish
Nothing else
16. What do you like about learning French/Spanish on Duolingo?
17. What do you want to see changed on Duolingo?
18. Did you have experience learning French/Spanish before using Duolingo? Yes/No
19. (If yes) How did you learn French/Spanish before using Duolingo? (Check all that apply)
Being around French/Spanish speakers
High school French/Spanish classes
College French/Spanish classes
Language apps
Internet-based materials such as podcasts and YouTube
Textbooks and other materials in print
Other (please specify)
20. Did you take French/Spanish classes during the time you used Duolingo? Yes/No
21. Did you use other programs or apps to learn French/Spanish during the time you used Duolingo? Yes/No

Note. *Any other Ls, whether L1 or L2.

L2 proficiency

Spanish elicited imitation test

(Solon et al., 2019)

Introduction

This language test will ask you to listen to several short audio files in Spanish and make a recording in response. (Please be patient as recordings may take time to load.)

Note that some of the items might be quite challenging. Please try to complete each of them to the best of your ability.

<Click here to start>

Instructions-1

You are going to hear several sentences in English (6 in total). After each sentence, there will be a short pause, followed by a tone sound {TONE}. Your task is to try to repeat exactly what you hear. You will have only one attempt to do so. You will be given sufficient time after the tone to repeat the sentence. Repeat as much as you can. Remember, don't start repeating the sentence until after you hear the tone sound {TONE}. Now let's begin.

<I'm ready>

Practice stimuli

1. We drove to the park.
2. I'll call her tomorrow night.
3. You can buy meat at the butcher shop.
4. My brother just bought a brand new computer.
5. Sometimes they take their dog for a walk in the park.
6. We're going to play volleyball at the gym that I told you about.

Instructions-2

Now, you are going to hear a number of sentences in Spanish (36 in total). Once again, after each sentence, there will be a short pause, followed by a tone sound {TONE}. Your task is to try to repeat exactly what you hear in Spanish. You will have only one attempt to do so. You will be given sufficient time after the tone to repeat the sentence. Repeat as much as you can. Remember, don't start repeating the sentence until after you hear the tone sound {TONE}. Now let's begin.

Main stimuli

1. Quiero cortarme el pelo. (7 syllables)
2. El libro está en la mesa. (7 syllables)
3. El carro lo tiene Pedro. (8 syllables)
4. Él se ducha cada mañana. (9 syllables)
5. ¿Qué dice usted que va a hacer hoy? (9 syllables)
6. Dudo que sepa manejar muy bien. (10 syllables)
7. Las calles de esta ciudad son muy anchas. (11 syllables)
8. Puede que llueva mañana todo el día. (12 syllables)
9. Las casas son muy bonitas pero caras. (12 syllables)
10. Me gustan las películas que acaban bien. (12 syllables)
11. El chico con el que yo salgo es español. (13 syllables)
12. Después de cenar me fui a dormir tranquilo. (13 syllables)
13. Quiero una casa en la que vivan mis animales. (14 syllables)
14. A ustedes les fascinan las fiestas grandiosas. (14 syllables)
15. Ella sólo bebe cerveza y no come nada. (15 syllables)
16. Me gustaría que el precio de las casas bajara. (15 syllables)
17. Cruza a la derecha y después sigue todo derecho. (15 syllables)

18. Ella ha terminado de pintar su apartamento. (14 syllables)
19. Me gustaría que empezara a hacer más calor pronto. (15 syllables)
20. El niño al que se le murió el gato está triste. (15 syllables)
21. Una amiga mía cuida a los niños de mi vecino. (16 syllables)
22. El gato que era negro fue perseguido por el perro. (16 syllables)
23. Antes de poder salir él tiene que limpiar su cuarto. (16 syllables)
24. La cantidad de personas que fuman ha disminuido. (16 syllables)
25. Después de llegar a casa del trabajo tomé la cena. (17 syllables)
26. El ladrón al que atrapó la policía era famoso. (17 syllables)
27. Le pedí a un amigo que me ayudara con la tarea. (17 syllables)
28. El examen no fue tan difícil como me habían dicho. (17 syllables)
29. ¿Serías tan amable de darme el libro que está en la mesa? (18 syllables)
30. Hay mucha gente que no toma nada para el desayuno. (18 syllables)
31. Son ellas las que acaban de decorar la sala de espera. (19 syllables)
32. ¿Sabe usted si el tren de las once y media ya ha salido de la estación? (20 syllables)
33. Nunca me divertí tanto como cuando fui a la pista de hielo. (20 syllables)
34. Cuanta más prisa tenía en su trabajo, menos calidad producía. (21 syllables)
35. Ellos lo organizaron el año pasado en la universidad cercana. (23 syllables)
36. Acabamos de volver del supermercado donde las ofertas eran muy interesantes. (27 syllables)

Thank you! Your responses have been recorded.

French elicited imitation test

(Gaillard & Tremblay, 2016)

Introduction

This language test will ask you to listen to several short audio files in French and make a recording in response. (Please be patient as recordings may take time to load.)

Note that some of the items might be quite challenging. Please try to complete each of them to the best of your ability.

<Click here to start>

Instructions-1

You are going to hear several sentences in English (6 in total). After each sentence, there will be a short pause, followed by a tone sound {TONE}. Your task is to try to repeat exactly what you hear. You will have only one attempt to do so. You will be given sufficient time after the tone to repeat the sentence. Repeat as much as you can. Remember, don't start repeating the sentence until after you hear the tone sound {TONE}. Now let's begin.

<I'm ready>

Practice stimuli

1. We drove to the park.
2. I'll call her tomorrow night.
3. You can buy meat at the butcher shop.
4. My brother just bought a brand new computer.
5. Sometimes they take their dog for a walk in the park.
6. We're going to play volleyball at the gym that I told you about.

Instructions-2

Now, you are going to hear a number of sentences in French (50 in total). Once again, after each sentence, there will be a short pause, followed by a tone sound {TONE}. Your task is to try to repeat exactly what you hear in French. You will have only one attempt to do so. You will be given sufficient time after the tone to repeat the sentence. Repeat as much as you can. Remember, don't start repeating the sentence until after you hear the tone sound {TONE}. Now let's begin.

Main stimuli

1. Dans cette grande ville, les rues sont larges.
2. Je doute qu'il sache si bien conduire.
3. Qu'est-ce que tu as dit que tu faisais?
4. Il est possible qu'il pleuve des cordes.
5. Les maisons sont très belles mais trop chères.
6. Le livre rouge n'était pas sur la table.
7. Ni lui ni moi ne les avions comprises!
8. Il prend une douche tous les matins à 7h00.
9. Je n'aime pas les films qui sont à l'eau de rose.
10. Après le déjeuner, as-tu fait une bonne sieste?
11. Tu aimes écouter la musique techno, n'est-ce pas ?
12. Est-ce que tu penses que je dois me faire couper les cheveux?
13. Traverse la rue au feu et puis continue tout droit!
14. Y-a-t-il beaucoup de gens qui ne mangent rien le matin?
15. On en avait une petite noire qui s'appelait minouche.
16. J'espère que le temps se réchauffera plus tôt cette année.
17. Le petit garçon dont le chaton est mort hier est triste.
18. Quand Sophie reçut sa collègue, elle lui proposa du thé.
19. Ce restaurant est censé avoir de la très bonne nourriture.
20. Je veux une belle et grande maison dans laquelle mes enfants puissent vivre.
21. La chatte que tu as nourrie hier était celle de ma voisine.
22. Le nombre de fumeuses en France ne cesse d'augmenter chaque année.
23. Gabriel, en épousant sa patronne, a fait d'une pierre deux coups.
24. N'êtes-vous pas fatigués après ce voyage en voiture de trois jours?
25. Nous aurions dû faire des réservations avant d'aller au théâtre.
26. Prenons deux semaines pour visiter New York pendant les vacances d'été!
27. Qu'allez-vous faire demain soir après lui avoir dit la vérité?
28. Est-ce qu'elle vient de finir de peindre l'intérieur de son appartement?
29. La personne avec qui je sortais n'avait pas un grand sens de l'humour.
30. Elle commande uniquement des plats de viande et ne mange jamais de légumes.
31. Vous pensez que le prix des maisons en ville va redevenir abordable?
32. Une bonne amie à moi s'occupe toujours des trois enfants de mon voisin.
33. Avant de pouvoir aller dehors, il doit finir de ranger sa chambre.
34. La police a arrêté le terrible voleur qui était grand et mince.
35. Auriez-vous la gentillesse de me passer le livre qui est sur la table ?
36. Elle a décidé de suivre des études d'arts plastiques à l'École des Beaux-Arts.
37. Dès que la présidente eut signé le document, son secrétaire l'emporta.

38. Excusez-moi, savez-vous si le train de 11h30 a déjà quitté la gare?
39. Je ne me suis jamais autant amusée que lorsque je suis allé à la patinoire.
40. Ce sont eux qui l'ont organisé l'an dernier à l'Université de l'Illinois.
41. Plus elle se dépêchait dans son travail, moins elle réalisait un travail de qualité.
42. Dès que l'on aura dîné, on regardera attentivement le documentaire sur France 3.
43. Ne penses-tu pas que les réalisatrices du film souhaiteraient lire les scénarios le plus tôt possible?
44. L'examen n'était pas aussi difficile que celui de Monsieur Durand en cours de littérature.
45. Laura et Julie, ce sont elles qui viennent de finir de décorer élégamment la chambre d'amis.
46. Il est possible que ses parents soient arrivés en France avant le début de la guerre d'Algérie.
47. On vient juste de rentrer du supermarché où les promotions étaient particulièrement intéressantes.
48. Les étudiants Laure et Stéphanie vont continuer à l'étudier à l'Université de Montréal.
49. Marie, prenez votre courage à deux mains et vous verrez que cet entretien passera comme une lettre à la poste!
50. Les étudiants sortant de l'université avec un Master en poche ont plus de chance de trouver un travail que les autres.

Thank you! Your responses have been recorded.

Notes on EIT administration and scoring

For the sake of comparability of the findings and the testing procedures, several modifications were made:

1. The original French stimuli were amplified in Audacity.
2. Sample items (practice stimuli) for both tests were taken from the Spanish version of the EIT and provided in English.
3. Both Spanish and French stimuli were presented in increasing length (not randomized as was the case with the original French stimuli).
4. An introduction for both French and Spanish EIT was added to explain the nature of the test.
5. The original instructions had to be slightly modified (due to the self-paced nature of both tests in the present study).
6. All instructions were recorded by a female speaker with a standard American dialect. After recording in Audacity, the peak amplitude was normalized to -1.0 dB (to help with the volume); a noise reduction for extraneous background noises was performed; and the beep sounds were added where indicated in the script.
7. For both French and Spanish EITs, a tone sound (.25s) from the original French EIT was used. A 3-second pause was inserted between the auditory sentence and the tone sound (as in the original French test). The total wait time between the auditory stimulus and the onset of sentence repetition was, therefore, 3.25s (as in the original French EIT study).
8. There were no breaks between the trials as in the Spanish EIT study.
9. As both EITs were administered as self-paced tests, the maximum recording time was estimated based on the formula by Solon et al. (2019; see supplementary materials) and set to 19s/19,000 ms (rounded based on the calculations below).
Spanish sentence #36 (Acabamos de volver del supermercado donde las ofertas eran muy interesantes) = 27 syllables, 6.248s (the longest sentence recorded by a native speaker)
27 syllables = 6.248s → native speaker time

7 syllables = $6.248 + 2 \rightarrow$ nonnative speaker time

27 syllables = $(6.248 + 2) + (20 \text{ syllables} * .5) = 8.248 + 10 = 18.248\text{s} \rightarrow$ max. recording time for nonnative speakers

French sentence #49 (Marie, prenez votre courage à deux mains et vous verrez que cet entretien passera comme une lettre à la poste!) = 28 syllables, 6.238s (the longest sentence recorded by a native speaker)

28 syllables = 6.238s \rightarrow native speaker time

7 syllables = $6.238 + 2 \rightarrow$ nonnative speaker time

28 syllables = $(6.238 + 2) + (21 \text{ syllables} * .5) = 8.238 + 10.5 = 18.738\text{s} \rightarrow$ max. recording time for nonnative speakers

10. Finally, for both French and Spanish tests, a rubric developed by Solon et al. (2019) was used.

Spanish C-test

(Riggs & Maimone, 2018)

Introduction

In this language test, you will be presented with short Spanish texts in which parts of words are deleted. The deletions correspond to the final portions of the words. Please do your best to fill in the missing part of the word.

Complete the words as accurately as possible, paying attention to the spelling and grammatical features like accents or agreement in gender and number.

You may put a zero in the blank if you do not know the answer and do not want to guess. There will be a total of 5 texts, each taking about 3 to 5 minutes to complete. Please try to finish each text in under 6 minutes.

Main part

Below you will be presented with short Spanish texts in which parts of words are deleted. The deletions correspond to the final portions of the words. Please do your best to fill in the missing part of the word. Complete the words as accurately as possible, paying attention to the spelling and grammatical features like accents or agreement in gender and number. You may put a zero in the blank if you do not know the answer and do not want to guess. There will be a total of 5 texts, each taking about 3 to 5 minutes to complete. Please try to finish each text in under 6 minutes.

Spanish accents (if you do not have a Spanish keyboard): á, é, í, ó, ú, ñ, ü.

Example

On Sunday, the weather was beautiful, and we went for a walk.

On Monday, it was raining, and we stay at home.

Text 1. *17 de agosto, Marbella*

Laura todavía está en la playa y yo estoy ya en la habitación del hotel. No quiero sa_____, hace mu_____ calor y no de_____ tomar m_____ el s_____. Todas l_____ mañanas va_____ a l_____ playa, a u_____ playa peq_____ pero m_____ bonita ce_____ del hotel. Allí no h_____ mucha ge_____. Después com_____ juntos en u_____ bar. Hay muc_____ en esa zo_____. ¡Cómo m_____ gusta l_____ comida de aq_____, sobre to_____ el pes_____! Por las

tar _____ vamos a _____ centro de Marbella. A veces volvemos muy pronto al hotel. Nos gusta mucho escuchar música o leer un buen libro.

Text 2. *Descripción de mi mamá Lauri*

Mi personaje favorito es mi mamá, su nombre es Lauriana de Jesús Pinta. Tiene cuar _____ y ocho añ _____ de ed _____ y tra _____ de prof _____ en el cen _____ educativo “Fiscal Francia” de ni _____. Ella vi _____ en Calvas c _____ mi pa _____ y mi her _____, a qui _____ quiero y apr _____ mucho. A mi madre la adm _____ porque e _____ una per _____ luchadora, respe _____, solidaria, soci _____, etc. Le gu _____ el dep _____, en espe _____ el basqu _____. Está siempre ale _____, es m _____ amorosa y comuni _____, y trata de darme y enseñarme lo mejor que tiene. Los momentos que he compartido con ella han sido inolvidables.

Text 3. *Así es el día a día de una cantante famosa*

Me levanto a las 8 y desayuno un panecillo con salmón ahumado. Mi hor _____ suele s _____ frenético, pe _____ mi ma _____ sabe l _____ que hay que ha _____ y pu _____ repararlo c _____ ella. Me gu _____ sorprenderme, a _____ que nu _____ miro mi calen _____ la no _____ anterior. Me enc _____ la sens _____ de levan _____ por l _____ mañana y te _____ que mi _____ por la ven _____ para desc _____ en q _____ ciudad es _____. Como norma _____ paso la may _____ del tiempo en el hotel, mi madre no se preocupa demasiado. Pocas veces discutimos, porque quiero muchísimo a mi madre.

Text 4. *Perfil de alimentación de los argentinos*

El último censo realizado en la República de Argentina contabiliza a su población en algo más de 40 millones de habitantes. El pa _____ produce una cant _____ suficiente pa _____ alimentar a 442 millones de pers _____, sin emb _____, por u _____ lado s _____ observan indiv _____ que pres _____ déficit de nutri _____ en s _____ alimentación, y por ot _____ lado, tam _____ excesos. A los argentinos les so _____ comida pero les fa _____ variedad. Hay homoge _____ en l _____ cocina y e _____ la me _____ de los argentinos. Se cons _____ pocos alim _____ de bu _____ calidad nutri _____, mientras que el exceso de con _____ de ot _____ agrega grasas de mala calidad, sodio y azúcares.

Text 5. *Un cubano en Kiev, recién llegado y sintiéndose agobiado*

Creo que este correo que les escribí a mis padres fue el más sentimental que he hecho en mi vida. Yo ha _____ salido ha _____ dos dí _____ de Cuba, y el _____ no hab _____ tenido noti _____ más, no sab _____ dónde est _____, ni có _____ iba, ni q _____ había si _____ de m _____. Yo me sen _____ muy tri _____, pero no po _____ decirles e _____. Empecé dici _____ que to _____ me iba bi _____ y que ha _____ salido estu _____ el vi _____, y s _____ querer m _____ lágrimas empe _____ a salir y no paraban de rodar por mis mejillas mientras escribía. Sabía que los estaba engañando, pero consideraba injusto preocuparlos; total, no resolvería nada.

French C-test

(Counsell, 2018)

Introduction

In this language test, you will be presented with short French texts in which parts of words are deleted. The deletions correspond to the final portions of the words. Please do your best to fill in the missing part of the word.

Complete the words as accurately as possible, paying attention to the spelling and grammatical features like accents or agreement in gender and number. Words with hyphens or apostrophes like “celui-ci” or “l’ami” count as one word.

You may put a zero in the blank if you do not know the answer and do not want to guess. There will be a total of 5 texts, each taking about 3 to 5 minutes to complete. Please try to finish each text in under 6 minutes.

Main Part

Below you will be presented with short French texts in which parts of words are deleted. The deletions correspond to the final portions of the words. Please do your best to fill in the missing part of the word. Complete the words as accurately as possible, paying attention to the spelling and grammatical features like accents or agreement in gender and number. Words with hyphens or apostrophes like “celui-ci” or “l’ami” count as one word. You may put a zero in the blank if you do not know the answer and do not want to guess. There will be a total of 5 texts, each taking about 3 to 5 minutes to complete. Please try to finish each text in under 6 minutes.

French accents (if you do not have a French keyboard): é, à, è, ù, â, ê, î, ô, û, ç, ë, ï, ü.

Example

On Sunday, the weather was beautiful, and we went for a walk.

On Monday, it was raining, and we stay at home.

Text 1. *Edda, la gourmande*

Je m'appelle Edda. J'ai 37 ans. Je su _____ née à Rome e _____ vis dep _____ neuf a _____ à Paris av _____ mon ma _____ italien. D'ori _____ italienne (m _____ père) e _____ française (m _____ mère), j _____ baigne dep _____ toute pet _____ dans c _____ deux cult _____. Gourmande e _____ curieuse dep _____ toujours*, c'e _____ tout nature _____ que j' _____ commencé à m _____ passionner po _____ la cui _____. Elle fa _____ vraiment par _____ de ma vie**. La magie est toujours là.

Note. *The answer key was slightly different from the test version and had “comme toujours” instead. **The answer key version (“la cuisin.e. Elle fait vraiment partie de ma vie”) was preferred to the test version (“la d _____ cuisine. El _____ fait vrai _____ partie de ma vie”).

Text 2. *La Martinique*

La Martinique est une île de 64 km de long qui s'étale sur à peine 20 km de large, et est dominée par la montagne Pelée qui culmine à 1397 m. Elle présente une grande diversité de paysages. Le S _____ est cons _____ de coll _____ à l _____ végétation p _____ abondante. L _____ Nord e _____ montagneux. L _____ plages surpr _____ par le _____ beauté e _____ leur incro _____ diversité, av _____ des coul _____ qui vo _____ du sa _____ blanc lumi _____ au no _____ volcanique. L' _____ est transp _____ et da _____ les fo _____

marins o_____ trouve beau_____ de pois_____ * colorés. Il y a les plages tranquilles du Sud-caraïbe, bordés de cocotiers, et celles plus tumultueuses, de la côte atlantique.

(Note. *Based on the author's suggestion in personal correspondence, "des bancs de poissons" in the answer key was replaced with "beau**coup** de poissons.")

Text 3. *Rester chez ses parents*

En Île-de-France, les jeunes restent un peu plus longtemps chez leurs parents que dans les autres régions françaises. Mais pourquoi cette tendance des jeunes à rester chez leurs parents s'accentue-t-elle en région parisienne ? Cette diffé_____ avec l_____ reste d_____ la France s'exp_____ essentiellement p_____ la prox_____ des unive_____. En ef_____, il y a beau_____ de facu_____ à Paris e_____ dans s_____ région. P_____ conséquent, l_____ jeunes d_____ l'Île-de-France q_____ font d_____ études univers_____ ne so_____ pas obl_____ de qui_____ le domi_____ familial. Ma_____ en prov_____, par con_____, les universités sont souvent éloignées du domicile des parents et les jeunes doivent quitter leur famille pour poursuivre leurs études.

Text 4. *La science-fiction*

La science n'a-t-elle pas de quoi considérer avec mépris les œuvres de science-fiction souvent basées sur des faits irréalistes et des connaissances approximatives ? La NASA a décidé d'en finir avec cette situation. La cél_____ agence spat_____ américaine vi_____ de lan_____ un nou_____ projet inti_____ « œuvres d_____ fictions insp_____ par l_____ NASA », e_____ collaboration av_____ un édi_____ de science-_____. L'idée e_____ de met_____ en con_____ des écri_____ avec l_____ scientifiques d_____ l'institution, af_____ qu'ils le_____ offrent u_____ expertise e_____ corrigent le_____ éventuelles err_____. L'objectif est simplement de produire des œuvres scientifiquement approuvées.

Text 5. *L'humour belge*

L'humour est belge, assurément ! Voilà une affirmation qui ne manquera pas de faire s'esclaffer la France entière. Mieux enc_____, elle se_____ l'objet d_____ quolibets div_____. C'est qu'_____ adorent ç_____, nos ch_____ voisins, ridic_____, critiquer, cho_____. À Paris rè_____ la gra_____ tyrannie d_____ persiflage. S_____ montrer d_____ et méc_____ est dev_____ gage d_____ réussite po_____ un humo_____. Goût d_____ scandale. Esca_____ de l_____ méchanceté. No_____ les Bel_____, avons l'hu_____ plus tendre, plus bon enfant. Rire ne signifie nullement gouailler, railler. Humour gentil, candide, humour belge.

The final message at the end of the pretest

This is the end of Part I of the study. As a reminder, please do your best to use Duolingo at least twice a week for a total of 26 weeks. Six months later, you will be invited to participate in Part II of the study. Thank you!

The final message at the end of the posttest

This is the end of Part II of the study. If you are eligible for compensation, you will be contacted by researchers within a month. Thank you!

Notes on C-test administration and scoring

1. Each test included five short texts with 25 blanks in each (i.e., 125 blanks in total).
2. Both tests were timed (30 minutes).
3. There was only one correct answer for each blank in the French C-test. In the Spanish C-test, 16 blanks had two correct answers.
4. For both tests, instructions and example sentences were provided in English (rather than in the target language as was in the original French C-test).
5. Small adjustments were made to the French C-test when three minor inconsistencies were revealed between the answer key and the test version.

Appendix B. Assumption checking

- RQ1a. The assumptions for paired samples *t*-tests by language were satisfactorily met (i.e., the dependent variable of proficiency gain scores was continuously scaled; the distribution of the differences in proficiency gain scores followed the normal curve and contained no extreme univariate outliers; the independent variable of test mode – oral vs. written – consisted of categorical data from two related groups).
- RQ1b. There were no major violations of assumptions for the two-sample *t*-tests. For the EIT gains in French vs. Spanish, the dependent variable was approximately normally distributed for each language group, but Levene's test was statistically significant, suggesting the lack of homogeneity of variances; nonetheless, the sample sizes for the two language groups were roughly equal, which does not require equal population variances. For the C-test gains in French vs. Spanish, the dependent variable was, again, approximately normally distributed for each language group, and Levene's test was not statistically significant (i.e., equal variances assumed). However, univariate outlier analysis revealed two extreme outliers ($|z| > .3.29$) on the EIT gains variable and six additional extreme outliers on the C-test gains variable. A close inspection of these scores did not indicate any red flags in participants' performance. Therefore, the analyses were conducted twice, with and without outliers, to allow for comparisons.
- RQ2a. To meet the assumptions for Pearson correlations, all extreme outliers ($|z| > .3.29$) were removed from the variables of interest (i.e., 6 from the Login and C-test gains variables, 4 from the Session, Minutes, Level reviews, Skill practice, and Tests variables, 2 from the Lessons and EIT gains variables, and 1 from the Stories variable) as they were found to affect the correlation estimates. The assumption of linearity was satisfied as indicated by the matrix scatterplot. Q-Q plots and histograms suggested minor deviations from normality. Therefore, bootstrapped Pearson correlations (based on 1,000 samples) with bias-corrected and accelerated confidence intervals were performed (final $N=233$).
- RQ2b. Prior to performing multiple regression analyses, all extreme univariate outliers were removed from the variables of interest. The strongest predictors were chosen based on correlational analyses (see RQ2a). However, in the model predicting EIT gains, the Sessions and Logins variables were highly correlated. To ensure the absence of multicollinearity, the Sessions variable was removed from the model because it had a weaker correlation with EIT gains than the Logins variable. Following the removal

of 10 multivariate outliers on three continuous predictor variables in the model predicting EIT gains and the removal of 19 multivariate outliers on two continuous predictor variables in the model predicting C-test gains, the assumptions of linearity; absence of multicollinearity; absence of autocorrelation; and normality, linearity, and homoscedasticity of residuals were met.

- RQ3. The assumptions for Pearson correlations between Duolingo app usage variables (i.e., frequency, duration, and intensity) and individual differences (i.e., L2 grit and motivation) were satisfied after removing extreme outliers ($|z| > .3.29$) from the variables of interest. The inspection of the matrix scatterplot supported the assumption of linearity. To account for occasional deviations from normality, which were indicated by Q-Q plots and histograms, bootstrapped Pearson correlations (based on 1,000 samples) with bias-corrected and accelerated confidence intervals were performed (final $N=260$).
- RQ4a. Concerning the assumptions for Pearson correlations, first, eight extreme outliers ($|z| > .3.29$) were removed from the gains variables (i.e., two from the EIT and six from the C-test gains), and four outliers were eliminated from L2 grit consistency of interest because they were found to affect the correlation values. A matrix scatterplot did not indicate any violations of linearity. To account for minor deviations from normality (as suggested by Q-Q plots and histograms), bootstrapped Pearson correlations (based on 1,000 samples) with bias-corrected and accelerated confidence intervals were performed (final $N=248$).
- RQ4b. To check the assumptions for multiple regression analyses, first, all extreme univariate outliers were removed from the variables of interest (see RQ4a). Five multivariate outliers on three continuous predictor variables were removed as well. All the assumptions of linearity, the absence of multicollinearity, the absence of autocorrelation, and normality, linearity, and homoscedasticity of residuals were met.

Address for correspondence

Luke Plonsky
Northern Arizona University
United States
lukeplonsky@gmail.com

Co-author information

Ekaterina Sudina
East Carolina University
sudinae22@ecu.edu

Publication history

Date received: 3 March 2023

Date accepted: 8 August 2023

Published online: 25 August 2023