The effects of frequency, duration, and intensity on L2 learning through Duolingo: A ‘natural’ experiment

Luke Plonsky, Northern Arizona University
Ekaterina Sudina, East Carolina University
# Table of Contents

Abstract ........................................................................................................................................... 2  
Introduction ...................................................................................................................................... 3  
Literature Review .............................................................................................................................. 4  
  Distributed Learning ....................................................................................................................... 4  
  Natural Experiment .......................................................................................................................... 5  
  Mobile-Assisted Language Learning ............................................................................................... 6  
  L2 Grit and Motivation ................................................................................................................... 7  
  The Duolingo Course Structure ..................................................................................................... 8  
The Present Study ............................................................................................................................. 8  
Method ............................................................................................................................................. 9  
  Participants ..................................................................................................................................... 9  
  Instruments and Materials ............................................................................................................. 12  
  Procedure ...................................................................................................................................... 14  
  Data Analysis ................................................................................................................................. 15  
Results ............................................................................................................................................. 17  
  Preliminary Analyses: Scale Data ................................................................................................. 17  
  Preliminary Analyses: Proficiency Data ....................................................................................... 18  
Discussion ....................................................................................................................................... 23  
  Conclusion ..................................................................................................................................... 25  
Acknowledgements ......................................................................................................................... 26  
References ....................................................................................................................................... 27  
Appendix A: The Full Instrument Used in the Study ...................................................................... 31
Abstract
Instructed second language (L2) research has frequently addressed the effects of ‘massed’ vs. ‘spaced’ practice (e.g., Nakata & Suzuki, 2019). 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.
**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 (i.e., “massed learning”) most effective, or is instruction more effective through more frequent (i.e., “spaced”) 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), 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 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). 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 (Dörnyei, 2009), 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.
Literture Review

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).

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 explains 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 repeated spaced rather than repeated massed practice. Nonetheless, 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.

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, 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 Kasprowicz 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 Kasprowicz 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 retention 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.

**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 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). 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. However, to our knowledge, no study to date has conducted a natural experiment to investigate the effectiveness of mobile-assisted language learning.

**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 / 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). One exception is Loewen et al.’s (2020) 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.

**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. (2020) raised concerns over participants’ attitudes to some app-related features (e.g., lack of interaction and limited variation in Duolingo tasks). 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.

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).
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).

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?
Method

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 for more). 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 the group are summarized in Table 1.

Table 1
Participant Characteristics

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>French (k = 139)</th>
<th>Spanish (k = 148)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>44.24</td>
<td>43.78</td>
</tr>
<tr>
<td>SD</td>
<td>14.53</td>
<td>14.40</td>
</tr>
<tr>
<td>Range</td>
<td>20−77</td>
<td>19−74</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>49</td>
<td>59</td>
</tr>
<tr>
<td>Female</td>
<td>90</td>
<td>85</td>
</tr>
<tr>
<td>Other</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>
### Native language(s)

<table>
<thead>
<tr>
<th>Language</th>
<th>121</th>
<th>87</th>
<th>137</th>
<th>93</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English + Other(s)</td>
<td>18</td>
<td>13</td>
<td>11</td>
<td>7</td>
</tr>
</tbody>
</table>

### Other languages spoken (excluding French/Spanish)

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>65</th>
<th>110</th>
<th>74</th>
</tr>
</thead>
<tbody>
<tr>
<td>English + Other(s)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>49</td>
<td>35</td>
<td>39</td>
<td>26</td>
</tr>
</tbody>
</table>

### Reasons for learning French/Spanish

<table>
<thead>
<tr>
<th>Reason</th>
<th>84</th>
<th>24</th>
<th>70</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>For travel</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>For school</td>
<td>4</td>
<td>1</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>For job-related purposes</td>
<td>18</td>
<td>5</td>
<td>40</td>
<td>10</td>
</tr>
<tr>
<td>For fun/leisure</td>
<td>121</td>
<td>34</td>
<td>112</td>
<td>28</td>
</tr>
<tr>
<td>For memory/brain acuteness</td>
<td>79</td>
<td>22</td>
<td>77</td>
<td>19</td>
</tr>
<tr>
<td>For social purposes</td>
<td>31</td>
<td>9</td>
<td>73</td>
<td>18</td>
</tr>
<tr>
<td>Other</td>
<td>16</td>
<td>5</td>
<td>21</td>
<td>5</td>
</tr>
</tbody>
</table>

### Other languages studied (excluding French/Spanish)

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>22</th>
<th>34</th>
<th>23</th>
</tr>
</thead>
<tbody>
<tr>
<td>English + Other(s)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>108</td>
<td>78</td>
<td>114</td>
<td>77</td>
</tr>
</tbody>
</table>

### Education

<table>
<thead>
<tr>
<th>Level</th>
<th>121</th>
<th>87</th>
<th>137</th>
<th>93</th>
</tr>
</thead>
<tbody>
<tr>
<td>Some high school</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High school</td>
<td>12</td>
<td>8</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>Associate’s degree</td>
<td>4</td>
<td>3</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>Bachelor’s degree</td>
<td>55</td>
<td>38</td>
<td>58</td>
<td>39</td>
</tr>
<tr>
<td>Master’s degree</td>
<td>40</td>
<td>28</td>
<td>55</td>
<td>37</td>
</tr>
<tr>
<td>Ph.D.</td>
<td>18</td>
<td>13</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Trade School</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Other</td>
<td>11</td>
<td>8</td>
<td>7</td>
<td>5</td>
</tr>
</tbody>
</table>

### Ethnicity

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>121</th>
<th>87</th>
<th>137</th>
<th>93</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asian</td>
<td>7</td>
<td>5</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>African American</td>
<td>4</td>
<td>3</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Caucasian</td>
<td>117</td>
<td>84</td>
<td>126</td>
<td>85</td>
</tr>
<tr>
<td>Latino or Hispanic</td>
<td>7</td>
<td>5</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Other</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Description</td>
<td>Mean 1</td>
<td>Mean 2</td>
<td>SD 1</td>
<td>SD 2</td>
</tr>
<tr>
<td>-----------------------------------------------------------------------------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>Self-rated level of French/Spanish when started using the app</td>
<td>2.38</td>
<td>2.52</td>
<td>1.46</td>
<td>1.53</td>
</tr>
<tr>
<td>Self-rated level of French/Spanish at pretest</td>
<td>4.18</td>
<td>4.29</td>
<td>1.31</td>
<td>1.56</td>
</tr>
<tr>
<td>Self-rated level of French/Spanish at posttest</td>
<td>4.43</td>
<td>4.44</td>
<td>1.37</td>
<td>1.43</td>
</tr>
<tr>
<td>Skills learned through Duolingo the most</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vocabulary</td>
<td>112</td>
<td>19</td>
<td>120</td>
<td>20</td>
</tr>
<tr>
<td>Grammar</td>
<td>87</td>
<td>15</td>
<td>102</td>
<td>17</td>
</tr>
<tr>
<td>Pronunciation</td>
<td>71</td>
<td>12</td>
<td>56</td>
<td>9</td>
</tr>
<tr>
<td>Listening</td>
<td>96</td>
<td>16</td>
<td>87</td>
<td>14</td>
</tr>
<tr>
<td>Speaking</td>
<td>55</td>
<td>9</td>
<td>56</td>
<td>9</td>
</tr>
<tr>
<td>Reading</td>
<td>104</td>
<td>18</td>
<td>114</td>
<td>19</td>
</tr>
<tr>
<td>Writing</td>
<td>69</td>
<td>12</td>
<td>80</td>
<td>13</td>
</tr>
<tr>
<td>Using Duolingo resources other than regular lessons</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stories</td>
<td>110</td>
<td>49</td>
<td>111</td>
<td>46</td>
</tr>
<tr>
<td>Podcasts</td>
<td>36</td>
<td>16</td>
<td>34</td>
<td>14</td>
</tr>
<tr>
<td>Tips</td>
<td>64</td>
<td>29</td>
<td>73</td>
<td>30</td>
</tr>
<tr>
<td>Nothing</td>
<td>14</td>
<td>6</td>
<td>25</td>
<td>10</td>
</tr>
<tr>
<td>Experience learning French/Spanish before using Duolingo</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>42</td>
<td>30</td>
<td>32</td>
<td>22</td>
</tr>
<tr>
<td>Yes</td>
<td>97</td>
<td>70</td>
<td>116</td>
<td>78</td>
</tr>
<tr>
<td>Ways of learning French/Spanish before using Duolingo</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Being around native speakers</td>
<td>17</td>
<td>9</td>
<td>40</td>
<td>18</td>
</tr>
<tr>
<td>High school classes</td>
<td>65</td>
<td>36</td>
<td>79</td>
<td>36</td>
</tr>
<tr>
<td>Language apps</td>
<td>41</td>
<td>23</td>
<td>46</td>
<td>21</td>
</tr>
<tr>
<td>Internet-based materials (e.g., podcasts, YouTube)</td>
<td>9</td>
<td>5</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>Textbooks and other materials in print</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>
### 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.

---

<table>
<thead>
<tr>
<th>Conversational language classes</th>
<th>25</th>
<th>14</th>
<th>22</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other</td>
<td>21</td>
<td>12</td>
<td>20</td>
<td>9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Taking French/Spanish classes in addition to Duolingo</th>
<th>134</th>
<th>96</th>
<th>140</th>
<th>95</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>5</td>
<td>4</td>
<td>8</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Using other programs/apps in addition to Duolingo</th>
<th>116</th>
<th>83</th>
<th>130</th>
<th>88</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>23</td>
<td>17</td>
<td>18</td>
<td>12</td>
</tr>
</tbody>
</table>

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

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 $\alpha$ 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 $\alpha$.

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 ($rs = .58-.67$ 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 $\alpha$ 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.
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 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.

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 x 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.

**Data Analysis**

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

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 review, 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. The 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.

Results

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(L2 grit) = .68; r(L2 perseverance of effort) = .69; r(L2 consistency of interest) = .58; r(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)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Pretest</th>
<th></th>
<th>Posttest</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>k</td>
<td>M</td>
<td>SD</td>
<td>α</td>
</tr>
<tr>
<td>L2 grit</td>
<td>7</td>
<td>3.96</td>
<td>.62</td>
<td>.84</td>
</tr>
<tr>
<td>L2 perseverance of effort</td>
<td>5</td>
<td>3.71</td>
<td>.73</td>
<td>.86</td>
</tr>
<tr>
<td>L2 consistency of interest</td>
<td>2</td>
<td>4.58</td>
<td>.63</td>
<td>.78a</td>
</tr>
<tr>
<td>L2 motivation</td>
<td>5</td>
<td>3.55</td>
<td>.77</td>
<td>.88</td>
</tr>
</tbody>
</table>

Note. k = number of items; M = mean; SD = standard deviation; 95% CI = 95% confidence intervals of coefficient alphas; L2 = second language. Spearman’s rho = .59a, .65b.
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 pretest 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).

Table 3
Descriptive Statistics for Proficiency Test Scores (N = 287)

<table>
<thead>
<tr>
<th>Mode</th>
<th>Oral (EIT)</th>
<th></th>
<th></th>
<th>Written (C-test)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Pre-</td>
<td>Post-</td>
<td>Pre-</td>
<td>Post-</td>
<td></td>
</tr>
<tr>
<td>Language</td>
<td></td>
<td>Fr</td>
<td>Sp</td>
<td>Fr</td>
<td>Sp</td>
<td>Fr</td>
</tr>
<tr>
<td>M</td>
<td>.12</td>
<td>.27</td>
<td>.17</td>
<td>.30</td>
<td>.28</td>
<td>.27</td>
</tr>
<tr>
<td>SD</td>
<td>.08</td>
<td>.15</td>
<td>.11</td>
<td>.16</td>
<td>.13</td>
<td>.12</td>
</tr>
<tr>
<td>Minimum</td>
<td>.005</td>
<td>.02</td>
<td>.03</td>
<td>.01</td>
<td>.02</td>
<td>.00</td>
</tr>
<tr>
<td>Maximum</td>
<td>.47</td>
<td>.77</td>
<td>.57</td>
<td>.84</td>
<td>.68</td>
<td>.68</td>
</tr>
<tr>
<td>Valid</td>
<td>132</td>
<td>138</td>
<td>129</td>
<td>140</td>
<td>139</td>
<td>148</td>
</tr>
<tr>
<td>Missing</td>
<td>7</td>
<td>10</td>
<td>10</td>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Note. M = mean; SD = standard deviation. Oral = elicited imitation test, written = C-test. Fr = French, Sp = Spanish.

Figure 1
Proficiency Test Gains in the Written vs. Oral Mode and by Language
RQ1: To what extent do learner gains differ when tested in the written vs. oral mode?

**RQ1a.** 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 $[−.08, .34]$.

**RQ1b.** 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.

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

**RQ2a.** As demonstrated in Table 4, learners’ self-determined exposure and behavior (i.e., frequency, duration, and intensity of the 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 4
*Pearson Correlations*\(^1\) Between Duolingo App Usage and Proficiency Gains (N = 233)

<table>
<thead>
<tr>
<th>Type</th>
<th>App usage</th>
<th>EIT gains</th>
<th>C-test gains</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Frequency</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logins*</td>
<td>(r = .14), BCa 95% CI [.03, .26]</td>
<td>(r = .03), BCa 95% CI [-.09, .17]</td>
<td></td>
</tr>
<tr>
<td>Sessions</td>
<td>(r = .12), BCa 95% CI [.00, .24]</td>
<td>(r = .06), BCa 95% CI [-.07, .19]</td>
<td></td>
</tr>
<tr>
<td><strong>Duration</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minutes</td>
<td>(r = .01), BCa 95% CI [-.16, .18]</td>
<td>(r = .20), BCa 95% CI [.05, .35]</td>
<td></td>
</tr>
<tr>
<td><strong>Intensity: Content and Curriculum</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lessons</td>
<td>(r = .21), BCa 95% CI [.07, .35]</td>
<td>(r = .26), BCa 95% CI [.13, .40]</td>
<td></td>
</tr>
<tr>
<td>Level review</td>
<td>(r = .21), BCa 95% CI [.07, .34]</td>
<td>(r = .11), BCa 95% CI [-.04, .27]</td>
<td></td>
</tr>
<tr>
<td>Skill practice</td>
<td>(r = -.06), BCa 95% CI [-.19, .07]</td>
<td>(r = .08), BCa 95% CI [-.04, .22]</td>
<td></td>
</tr>
<tr>
<td>Stories</td>
<td>(r = -.02), BCa 95% CI [-.16, .14]</td>
<td>(r = .07), BCa 95% CI [-.04, .18]</td>
<td></td>
</tr>
<tr>
<td>Tests</td>
<td>(r = .09), BCa 95% CI [-.02, .25]</td>
<td>(r = .09), BCa 95% CI [-.06, .22]</td>
<td></td>
</tr>
</tbody>
</table>

**Notes.** \(^1\)Bias-corrected and accelerated 95% confidence intervals for correlation coefficients.

*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.

### RQ2b.
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 Review’ emerged as the only meaningful positive predictor (see Table 5\(^1\)).

### Table 5
Regression Analysis Summary for Variables Predicting EIT Gains

<table>
<thead>
<tr>
<th>Variable</th>
<th>(B)</th>
<th>SE</th>
<th>(\beta)</th>
<th>(t)</th>
<th>(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target language</td>
<td>(-.008)</td>
<td>.006</td>
<td>-.095</td>
<td>-1.49</td>
<td>.14</td>
</tr>
<tr>
<td>Logins</td>
<td>(.000)</td>
<td>.000</td>
<td>.059</td>
<td>.887</td>
<td>.38</td>
</tr>
<tr>
<td>Lessons</td>
<td>(8.820\times 10^{-6})</td>
<td>.000</td>
<td>.051</td>
<td>6.58</td>
<td>.51</td>
</tr>
<tr>
<td>Level review</td>
<td>(.000)</td>
<td>.000</td>
<td>.158</td>
<td>2.052</td>
<td>.04</td>
</tr>
</tbody>
</table>

**Note.** N = 243. Overall model: \(R^2 = .06\), adjusted \(R^2 = .04\).

\(^1\) Of note, the Sessions variable was excluded from the model due to a large correlation with the Logins variable.
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, R2 = .06, adjusted R2 = .05. ‘Lessons’ emerged as the only meaningful positive predictor (see Table 6).

Table 6
Regression Analysis Summary for Variables Predicting C-test Gains

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

Note. N = 261. Overall model: R2 = .06, adjusted R2 = .05.

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 7, 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 7
Pearson Correlations1 Between Duolingo App Usage and Individual Differences (N = 260)

<table>
<thead>
<tr>
<th>Type</th>
<th>App usage</th>
<th>L2 Perseverance of Effort</th>
<th>L2 Consistency of Interest</th>
<th>L2 Motivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>Logins</td>
<td>r = .16 [.06, .27]1</td>
<td>r = .08 [-.05, .20]</td>
<td>r = .20 [.08, .31]</td>
</tr>
<tr>
<td>Frequency</td>
<td>Sessions</td>
<td>r = .15 [.04, .26]</td>
<td>r = .12 [-.02, .23]</td>
<td>r = .20 [.07, .31]</td>
</tr>
<tr>
<td>Duration</td>
<td>Minutes</td>
<td>r = .18 [.05, .30]</td>
<td>r = .13 [.02, .22]</td>
<td>r = .24 [.12, .35]</td>
</tr>
</tbody>
</table>
RQ4: To what extent are L2 grit and motivation associated with gains in L2 Spanish and French?

RQ4a. 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; $r = .12$ ($r_{corrected} = .22$), 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; $r = .16$ ($r_{corrected} = .19$), BCa 95% CI $[.04, .28]$ between C-test gains and L2 perseverance of effort (note that the confidence interval does not cross zero); $r = .09$ ($r_{corrected} = .17$), 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).

RQ4b. 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.
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-internal and learner-external) 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 spaced vs. massed learning 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 (e.g., 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. 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. In light of this finding, one suggestion for efficient use of Duolingo is moving forward and learning new content. Regardless of the source, 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. In short, less gains necessarily means less for the predictor or independent variables to explain.
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. The 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.

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 the strongest associations with in-app exposure. However, we also observed meaningful and statistically significant correlations between L2 perseverance of effort and several in-app behaviors.

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 R2 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.

**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 but was carried out independently by the authors. We are very grateful to the Learning Science Lab 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, Jay Rosok, Bozena Pajak, 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


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

<table>
<thead>
<tr>
<th>Not like me at all</th>
<th>Not much like me</th>
<th>Somewhat like me</th>
<th>Mostly like me</th>
<th>Very much like me</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

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.

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 llvea 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 sensé 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. Elle commande uniquement des plats de viande et ne mange jamais de légumes.
30. Vous pensez que le prix des maisons en ville va redevien abordable?
31. Une bonne amie à moi s’occupe toujours des trois enfants de mon voisin.
32. Avant de pouvoir aller dehors, il doit finir de ranger sa chambre.
33. La présidente eut signé le document, son secrétaire l’emporta.
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’Ecole 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 étudiantes 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/19,000ms (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)

\[ 27 \text{ syllables} = 6.248 \text{s} \times \text{native speaker time} \]
\[ 7 \text{ syllables} = 6.248 + 2 \times \text{nonnative speaker time} \]
\[ 27 \text{ syllables} = (6.248 + 2) + (20 \text{ syllables} \times 0.5) = 8.248 + 10 = 18.248 \text{s} \times \text{max. recording time for nonnative speakers} \]

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)

\[ 28 \text{ syllables} = 6.238 \text{ s} \times \text{native speaker time} \]
\[ 7 \text{ syllables} = 6.238 + 2 \times \text{nonnative speaker time} \]
\[ 28 \text{ syllables} = (6.238 + 2) + (21 \text{ syllables} \times 0.5) = 8.238 + 10.5 = 18.738 \text{s} \times \text{max. recording time for nonnative speakers} \]

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______ repasarlo 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______ mías, 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, ü, è, ë, ï, 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 suis née à Rome envoyées depuis neuf ans à Paris avec mon mari italien. D'origine italienne (mon père) et française (ma mère), je baigne depuis toute petite dans deux cultures. Gourmande et curieuse depuis toujours*, c’est tout naturel que j’aies commencé à m’passionner pour la cuisine. Elle fait vraiment partie de ma vie**. La magie est toujours là.

*The answer key was slightly different from the test version and had “comme toujours” instead. **The answer key version (“la cuisine. 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’e_______ 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 “beaucoup 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’idé_______ de met_______ en con_______ des écri_______ avec l_______ scientifiques d_______ l’institution, af_______ qu’ils le_______ offrent u_______ expertise e_______ corrignent
le_____ éventuelles erreurs. L’objectif est simplement de produire des œuvres scientifiquement approuvées.

**Text 5: L’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.