- Roumen Vesselinov, PhD
- John Grego, PhD
- Mila Tasseva-Kurktchieva, PhD
- Nasrin Sedaghatgoftar, PhD

THE BUSUU EFFICACY STUDY

March
2021

KEYWORDS

Busuu, efficacy, Language Learning Apps, Computer Assisted Language Learning (CALL), Mobile Assisted Language Learning (MALL), Educational Technology

- University of Maryland
- University of South Carolina
- University of South Carolina
- Kharazmi University

◆ Corresponding author: rvesselinov@som.umaryland.edu

The Busuu Efficacy Study 2021

Executive Summary

This study is based on a random sample of 114 Busuu users, with 35% beginner/novice users (placed in 1st college semester of Spanish) and 65% intermediate users (semester 2,3,4).

All participants took at the beginning of the study 2 language tests: one for reading/grammar proficiency (WebCAPE) and one for oral proficiency (TNT). All participants used Busuu to study Spanish for 2 months and took the same tests again at the end of the study.

MAIN RESULTS

Overall Language Proficiency Improvement

1. 94% improved in at least one of the areas: reading/grammar or oral proficiency (95%CI 88-98)
2. 80% improved their reading/grammar proficiency (95%CI 71-86)
3. 71% improved their oral proficiency (95%CI 62-78)
4. 56% improved both their reading/grammar and oral proficiency (95%CI 46-65)
5. 53% moved up at least 1 semester in their college placement (95%CI 44-62)

Language Proficiency Improvement for Beginner and Intermediate Users

1. 100% of the beginner users improved in at least one of the areas: reading/grammar or oral proficiency, compared to 82% of the intermediate users.
2. 92% of the beginner users improved their reading/grammar proficiency, compared to 73% of the intermediate users.
3. 75% of the beginner users improved their oral proficiency, compared to 68% of the intermediate users.
4. 67% of the beginner users improved both their reading/grammar and oral proficiency, compared to 50% of the intermediate users.
5. 77% of the beginner users moved up at least 1 semester in their college placement, compared to 41% of the intermediate users.

Busuu Efficacy: proficiency gain per one hour of study.

- Reading/Grammar Proficiency Efficacy
  1. Busuu users on average gain 5.8 WebCAPE points per one hour of study
  2. Busuu users will need on average 13 hours of study in a two-month period to move up one college semester (from 2nd to 3rd), (95%CI 10-23).
- Oral Proficiency Efficacy
  1. Busuu users on average gain 0.036 TNT points per one hour of study.
  2. Busuu users will need on average 28 hours of study in a two-month period to increase their oral proficiency by one full level (95%CI 20-50).

Results Based on TNT Estimates of CEFR and ACTFL

- Reading/Grammar Proficiency
  1. Based on CEFR estimate, 57% improved their reading/grammar proficiency (95%CI 47-65)
- Oral Proficiency
  1. Based on CEFR estimate, 46% improved their oral proficiency (95%CI 37-55).
  2. Based on ACTFL estimate, 48% improved their oral proficiency (95%CI 39-57).

Main Efficacy Factors

- Existing Language Skills
  1. Language aptitude
  2. Existing second language knowledge
  3. Initial language proficiency
  4. Motivation
- Structure and Length of the Study Time
  1. Total study time (app and Live lessons)
  2. Number of Live lessons
  3. Percent of time with Live lessons
  4. Study time working with Busuu app

Table of contents

Page no

1. Introduction

1.1 The Busuu Teaching Approach

Busuu comprises a website (busuu.com), an iOS and an Android mobile application (app). As of November 2020, Busuu claims to have over 115 million registered users in 190 countries, with an average of 30,000 new users registering each day. Busuu offers courses in 12 languages, of which English is the most popular, studied by 46% of all users and Spanish the second most popular, studied by 10% of users. Although aimed primarily at adult learners, Busuu has users from age 14 upwards. Busuu operates a freemium business model in which certain content and features are available to all registered users but additional content and features are accessed via a premium subscription, which costs around 10 Euros per month depending on the market. For this study, all participants were provided with full access to the Premium version of Busuu.

The Busuu Spanish course covers Common European Framework of Reference for Languages (CEFR) levels A1 to B2 and is translated into 15 different interface languages (Arabic, Chinese, English, French, German, Indonesian, Italian, Japanese, Korean, Polish, Portuguese, Russian, Turkish, Vietnamese). The interface language allows a learner to access instructions and explanatory content in their native language. The Spanish syllabus focuses on teaching communicative skills, with each lesson or group of lessons presenting functional language and building towards a Conversation exercise in which students write or record a short response to a prompt (e.g. Introduce yourself to the Busuu community), and share these responses with native or advanced-level speakers in the community to receive peer feedback. All users are requested to give feedback to learners of the languages they speak fluently but doing so is voluntary and around 69% of Spanish learners choose not to. Unlike many other language learning apps, Busuu provides more than just vocabulary memorization exercises. Lessons focus on different skills, including grammar, listening, and reading. Pronunciation is reinforced through speech recognition exercises and learners are provided with opportunities to practice both receptive and productive skills.

In addition to the general Spanish learning content, Busuu provides special courses including Spanish for Travel, Learn Spanish with El País and a series of Spanish language podcasts. Learners can also use the adaptive Vocabulary and Grammar Review features which enable them to memorize new phrases and practice, for example, conjugating verbs in a particular tense. Each Review session is personalized to the learner and algorithmically generates a unique set of exercises based on how well the learner has performed on these topics in the past.

Since July 2020, one to one online lessons have been available to all Busuu users. This feature is called Busuu Live. These are pre-bookable 30-minute lessons with professional language teachers, who provide classes tailored to the specific needs of each student. Teachers provide their own curriculum but are given information on the students' progress through the Busuu app, so are able to tailor lessons to match the students' level and attainment.

In summary, the Busuu teaching methodology comprises three elements: self-study using the Busuu app, practice with Spanish speakers via the Busuu community and lessons with professional Spanish teachers through Busuu Live.

1.2 The Current Study

The current study seeks to evaluate the efficacy of Busuu's Complete Spanish course and Busuu Live lessons with registered Busuu learners. The study uses an experimental, pre-test/post-test design. The research design is the same as in the previous dozen efficacy studies by the research team (Vesselinov, Grego, et al., 2009-2020). We use a random selection of the study sample, test the participants in the beginning of the study and at the end and measure the difference between the two sets of tests. Only the use of Busuu app and Live lessons was allowed during the two-month study.

1.3 Research Questions

The current study will add to the current pool of efficacy research on language learning apps, particularly of Busuu with updates to its curriculum and features and the addition of Live lessons. It will answer the following questions:

1. What is the efficacy of Busuu's Complete Spanish course with Live lessons?

2. What factors affect the efficacy?

The pool of potential factors includes motivation, language aptitude, native language, previous experience of learning foreign language, amount of study time, balance of time spent on the app and in Live lessons, etc.

The main measures for success are:

- Reading/grammar proficiency gain – the difference between the final and initial reading/grammar test score.
- Oral proficiency gain – the difference between the final and initial oral proficiency test score.
- Reading/grammar efficacy – the reading/grammar gain divided by the individual study time.
- Oral proficiency efficacy – the oral proficiency gain divided by the individual study time.

2. Literature Review

2.1 Theory on Mobile-Assisted Language Learning (MALL)

MALL is an extension of computer-assisted language learning (CALL). Chinnery (2006) traces the use of telephones in language learning back to the late 1980s, and the first use of mobile telephones for language learning to the Stanford Learning Lab, which incorporated them into a Spanish learning program in 2001. While early MALL technologies utilized PDAs and MP3s as support tools for the delivery of blended language learning programs, the rapid development of mobile technology has enabled smartphones to become providers of fully digital language training. With a smartphone or tablet, language learners can now access resources via marketplaces for online language tutors and self-study language learning applications (LLAs), "anytime, anywhere." Kukulska-Hume and Traxler state that "mobile learning is also personal learning, which could be remote and individual, or social and collaborative (2005, pp. 30-31)."

Burston (2014) identifies some of the current pedagogical challenges of MALL, particularly the dominance of teacher-centered approaches where MALL devices are seen purely as a method of delivering content to the learner. Better use of MALL should aim to provide opportunities for a more learner-centered, communicative pedagogy in which the technology is used to facilitate interaction and communication. It is, however, worth pointing out that this review surveys a wide range of MALL technologies such as MP3 and text messages. These technologies might have been popular at the time of this review, but they do not reflect the rapid development of mobile technologies in the past decade (see Heil et al., 2016 for a review on trends in this field). With different technical configurations, advanced machine learning and added multimedia functionality, LLAs are fundamentally different from their early counterparts. It is therefore crucial to see that many of the challenges Burston (2014) outlines might not be applicable to LLAs.

2.2 Efficacy in Second Language Acquisition

Both effectiveness and efficacy studies evaluate the impact of a treatment (Institute of Education Sciences & Foundation, 2013). Efficacy studies examine the impact of an intervention under ideal and controlled conditions, whereas effectiveness studies do so in "real-world" conditions. The current study uses an efficacy study approach because the learning process of participants is monitored by the research team.

In second language acquisition, a variety of individual and group differences influence learning results. Among them, language aptitude and motivation have been considered the two most impactful; language aptitude is a representative cognitive variable of the learner and motivation is an affective one (see Dörnyei, 2010 for a summary and history on these two constructs). Measuring both variables can provide a rather comprehensive insight into the learners' contribution in the learning process (Dörnyei, 2010).

2.2.1 Language Aptitude

Language aptitude refers to an individual's ability in learning languages, which can predict the speed and ease thereof (Cook, 2008; Macaro et al., 2010). Various language aptitude tests have been developed to measure this construct, which is multifaceted and found to be associated with other variables such as working memory capacity and language learning strategies (see Sparks & Ganschow, 2001 for a review). For example, the influential Modern Language Aptitude Test (MLAT) measures the learner's cognitive capacities in phonetic coding, grammatical sensitivity, rote learning of materials, and inductive language learning (Carroll & Sapon, 1959). A meta-analysis surveyed 33 studies on language aptitude and second language grammar acquisition over the span of five years and found language aptitude is moderately associated with second language (L2) grammar learning (r = .31) (Li, 2015). The results also show that language aptitude measures have more implications in early stages of foreign language learning (Li, 2015). Although MLAT has produced excellent predictions in the field, its outdated nature warrants reconsideration of the instrument (Sparks & Ganschow, 2001).

In an empirical study, Safar and Kormos (2008) examine the conceptualization and predictive strength of the Hungarian version of MLAT (HUNLAT) with 61 students in communicative classrooms. The results show that the HUNLAT scores only correlate weakly with learning outcomes in a communicative setting. Sedaghatgoftar et al. (2019) developed an updated language aptitude test – the Second Language Pragmatics Aptitude Test (SLPAT) – which demonstrated high construct validity for three factors: memory for pragmatic rule learning, mind-reading (ability to interpret social and emotional cues) from films and mind-reading from voices. The choice to use Sedaghatgoftar et al.'s (2019) language aptitude test is thus relevant and important in the current study for its focus on evaluating pragmatic language usage, which is in line with Busuu's communicative teaching focus.

2.2.2 Motivation

Motivation is a dynamic and complex affective construct that seeks to capture the reasons behind the initiation and sustenance of a learner's language learning process. In recent decades, research into motivation has become a dynamic area of study in the field of second language acquisition (SLA) (see Macaro et al., 2010 for an overview). Since its conception, the L2 motivation self system (L2MSS) has rapidly become the dominant model for research into L2 motivation (Dörnyei & Ryan, 2015; Lamb, 2017). The L2MSS proposes that language learners are more or less motivated depending on how strongly they are able to envision their future 'second language selves'.

Validation studies across a variety of cultural contexts have found that the L2MSS accurately predicts motivated learning behavior: for example, Kormos, Kiddle and Csizér (2011) in Chile, Taguchi, Magid and Papi (2009) in Iran, China and Japan; Kormos and Csizér (2008) in Hungary. These studies generally find the ideal L2 self to be a strong predictor of intention to persist with language learning, explaining 40% or more of the variance in criterion measures (Dörnyei & Ryan, 2015).

Kong et al. (2018) created a 33-question version of the L2MSS which included additional sections on 'international posture', 'intended effort' and 'competitiveness' which they used to assess 1296 Korean college age language learners. Their results further validated the L2MSS, indicating that the L2 learning environment was the best predictor of motivated attitudes, followed by the ideal L2 self.

2.2.3 Crosslinguistic influence

Crosslinguistic influence (CLI) refers to the impact of a language on the acquisition of another (Alonso, 2019). Carvalho and da Silva (2006) examined Spanish-English bilingual speakers learning Portuguese as a third language (L3) and found that linguistic similarity is more important than order of acquisition. Regardless of their first language, these bilingual speakers used Spanish to scaffold their Portuguese learning.

2.3 Previous Efficacy Studies

As commercial language learning apps have become more popular, a number of research studies have investigated their effectiveness and efficacy. For instance, Jiang et al. (2020) compared learning results of French and Spanish learners on Duolingo to those of university students from two previous studies. Although it is unsafe to assume that these cohorts could be compared directly due to differences in time and space of the intervention and the demographics of the participants, their results demonstrate that LLAs can be powerful in training reading and listening skills. On a broader level, Huang (2020) surveyed 32 qualified empirical studies on the effectiveness of LLAs in an unpublished systematic review. The results show mostly positive outcomes of learning a foreign language through apps. Although these studies investigated all aspects of language learning, vocabulary acquisition was the subject of interest in most studies. Studies concerned with communicative language skills like listening and speaking are small in number. At the same time, most empirical studies on this topic should improve on their study design and reporting practices. These insights corroborate those in Burston's (2014) meta-analysis on MALL devices.

Heil et al.'s (2016) review points out that the majority of MALL devices focus on training isolated vocabulary using rote memorization. Few applications provide a fully-fledged language learning package. In its design Busuu intends to cater to different language skills by placing a heavy focus on communicative teaching and grammar training. A number of previous studies have examined Busuu's efficacy and usage. Vesselinov & Grego (2016) evaluated an older version of the Busuu app with 144 learners of Spanish. The results show that Busuu was facilitative in advancing grammar knowledge as well as speaking and reading skills. Rosell-Aguilar (2018), in a survey of 4095 adult Busuu users around the world who spoke English and Spanish, found that 83% of participants agreed or strongly agreed that Busuu had helped them to develop their language skills. Over a third (36%) used Busuu as their only source of language learning, 40% used only apps and digital resources for language learning, and just 24% were taking part in any formal language learning program. Rosell-Aguilar's research suggests that a large group of adults globally are meeting their language learning needs only with an app, indicating the importance of further efficacy research into LLAs.

3. Methodology

3.1 Statistical Methodology

In the analysis for this study, we used some descriptive measures of central tendency, including means, standard deviation (SD), median (Me), first (Q1) and third (Q3) quartiles. We built standard 95% Confidence Intervals (CI) for means and proportions, using the Agresti-Coull correction (Agresti & Coull, 1998) for the latter. We used a Chi-square test and Fisher's exact test to relate two categorical variables, and ANOVA for continuous measures with a categorical factor. We also evaluated the relationship between two continuous variables with the Pearson correlation coefficient (r).

The most important consideration in this analysis is the non-linearity of the effects. We have discovered this phenomenon in our previous efficacy studies (Vesselinov, Grego, et al., 2009-2020). Most mainstream statistical methods and models like OLS regression and correlation assume that two variables have a more or less linear relationship. For example, we could reasonably expect that motivation level affects the improvement in language abilities. After building a statistical model, usually we are able to say that on average 1% (or point) increase in motivation leads to X% (or points) improvement in oral proficiency. But this is not the case. We have discovered that in most cases there is a threshold for the effect to appear. For example, as we will see in the analysis below, motivation has a positive effect on the results, but users may have to reach a certain very high level of motivation before the effect is visible.

In order to discover the existence and estimate these thresholds we used recursive-partitioning methods, specifically the Classification and Regression Tree (CART) models (Breiman et al., 1984 and Hastie et al., 2009).

We built CART multivariate models that relied on a pool of all potential factors that included all relevant variables available in this study. CART uses recursive partitioning methods to discover the most important factors and combination of factors that affect the outcome. It builds classification tree models for binary outcomes (e.g., any oral proficiency gain, Yes/No) and regression tree models for continuous outcomes (e.g., size of the oral proficiency gain in TNT points). For the purposes of this analysis, CART provides two important pieces of information. First, it gives the Variable Importance Measure (VIM) for all factors. VIM assigns a score of 100 to the most important variable in the model, followed by the second most important variable and so on. Second, CART discovers different groups of users and paths to success (e.g., improved oral proficiency).

We used the Receiver Operating Characteristic Area Under the Curve (ROC AUC) to evaluate the predictive quality of the CART models for the binary outcomes. Usually, a model with ROC AUC of 0.7 (or 70%) or above is considered a model with good predictive quality.

3.2 Participants

The study started in September of 2020, with Busuu sending notifications to recently registered users in the app informing them of the study. The inclusion criteria included Spanish as the learning language and the geographical location: US, UK, Australia, New Zealand, Canada, Brazil, or Portugal. The eligibility criteria also included age of at least 18 years old and the agreement to participate in the study. The participants agreed to use Busuu for two months to study Spanish with at least 2 hours of study time. For the oral proficiency evaluation, we recommended at least 8 hours of study. The use of other language apps or external language courses was prohibited.

The selection procedure is presented in the Appendix, Figure A1. We randomly selected 150 participants out of the 747 eligible respondents and invited them to take the initial language tests. 141 of them completed the initial language tests and this was our initial random sample. At the end of the study, we invited all participants with at least 2 hours of study to take the final tests. 119 of the participants completed the tests and this was our study sample. Of the initial sample of 141 participants, 22 did not complete the study because they did not have 2 hours of study or did not complete the final tests. The drop-out rate for this study was 15.6% (22 out of 141).

We compared the drop out group with the final sample on age, gender, education, and initial language tests results and there was no statistically significant difference between the two groups (p<0.01).

A small number of the participants (n=5) in the final study sample reported using other language apps on a regular basis and they were excluded from the final analysis. The final analysis sample size was 114 with 28 participants from Brazil and the rest (n=86) from English language speaking countries.

The native language for all Brazilian participants was Portuguese. For the rest, 88.4% (n=76) were native English speakers and 11.6% (n=10) had other native languages. The other languages included Bulgarian, Chinese, French, Italian, Japanese, Lithuanian, Polish, Punjabi, Russian and Tagalog.

The average age was 38.2 years (SD=12.4), and the age range was from 19 years to 76 years. 46.8% of the participants were female. The sample was well educated with 82.3% having a bachelor's degree or higher. Most of the participants (76.1%) worked full-time or part-time, with 12.4% unemployed and 11.4% retired, homemakers, or other.

The majority of participants (59.6%) studied Spanish for personal reasons, 23.7% for travel, and 14.9% for business or work. A quarter of the participants (22.8%) had a close friend or spouse who spoke Spanish and a small portion (6.1%) had parents or grandparents who spoke Spanish.

About 42% of the participants declared that they knew a second language, other than their native one. Some of the participants had extensive exposure to foreign languages, about 17% of them had lived in a foreign language-speaking country for more than 6 months and 9.6% had grown-up in a multilingual family.

For the Live lessons all participants had to have access to a bigger screen (desktop, laptop or tablet). About 83% of the users self-reported using smartphones to study Spanish with the Busuu app. The majority of the participants (78%) had used other language apps in the past.

About 92% of the participants defined themselves as beginning users of Spanish in the initial entry survey but the initial WebCAPE college placement test placed only 34.5% of them in the first college semester level of Spanish. About 32.7% were placed in the second semester, 17.7% in the third semester and 15% in fourth the college semester level of Spanish. This indicates that we cannot rely on the participants' evaluation of their own initial language level as it is routinely done in studies (Jiang et al., 2020).

Participants were incentivized to take part in the study by being offered up to 16 free Live Spanish lessons plus a free Busuu Premium account for one year for themselves and a friend of theirs.

3.3 Data Collection Procedure

In September of 2020, Busuu sent notifications for the study with a link to an online entry survey. The research team collected the surveys, created the pool of eligible participants and randomly selected 150 participants for the study. All participants took the two online tests managed by Emmersion Learning, Inc., then studied Spanish for 8 weeks and had access to 2 free Live lessons a week with professional Spanish language teachers and took the same tests again at the end of the study. The study time was extracted from Busuu servers on a weekly basis. Every week the research team sent emails to inform the participants of their study time and the number of Live lessons taken.

3.4 Study Instruments

3.4.1 TrueNorth Test (TNT)

This is a newly developed (2019) online oral proficiency test[1] based on elicited imitation and free speech as a testing method in which participants hear an utterance in the target language and are prompted to repeat the utterance as accurately as possible. TNT has good psychometric properties and reliability with Cronbach's Alpha=0.932 (Habing et al., 2020).

TNT gives an incremental oral proficiency score from 0.0 to 10.0 with zero being the lowest level and 10 – the highest. There are total of 100 possible TNT levels (values). TNT also provides estimates of the American Council for Teaching Foreign Languages (ACTFL)[2] level and the CEFR[3] level.

TNT estimation of ACTFL oral proficiency is denoted as TNT_ACTFL in this study. ACTFL has developed a proficiency scale to assess foreign language abilities. This scale includes four main groups (Novice, Intermediate, Advanced, and Superior), with the first three divided into levels as follows: Novice: 1. Low 2. Mid 3. High; Intermediate: 4. Low 5. Mid 6. High; Advanced: 7. Low 8. Mid 9. High, and 10. Superior.

TNT estimation of CEFR oral proficiency is denoted as TNT_CEFR in this study. CEFR is designed as a global standard for describing language proficiency. It has six levels, A1-A2 for beginner, B1-B2 for intermediate, and C1-C2 for proficient.

3.4.2 Web based Computer Adaptive Placement Exam (WebCAPE)

WebCAPE is an established university placement test and is offered in English, Spanish, French, German, Russian and Chinese. It was created by Brigham Young University and was acquired by Emmersion Learning. The WebCAPE test is used as a reading and grammar proficiency evaluation tool for placement in college-level language courses.

[1] https://emmersion.ai/products/truenorth/ .

[2] https://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012

[3] https://www.coe.int/en/web/common-european-framework-reference-languages

WebCAPE has high reliability (test-retest) value of 0.86[4] . The test is adaptive so the time for taking the test varies with an average time of 20-25 minutes. WebCAPE generates a score (in points) from 0 to 1000 (app.) and based on that score places the students in different college semesters; first (0-269), second (270-345), third (346-428), and fourth semester with more than 428 WebCAPE points. WebCAPE also generates an estimate of CEFR levels (denoted as WC_CEFR in this study) from A1 to C2.

Both TNT and WebCAPE place test-takers into an American college semester based on their test scores. One college semester is typically 15 to 18 weeks. At Brigham Young University, one semester is about 15 weeks (BYU 2021 Academic Calendar 2021). First and second semester Spanish courses comprise five class hours and two lab hours per week; third, fourth, and fifth semester courses consist of five class hours and one lab hour per week (BYU Class Search 2021). As a result, the contact hours in one college semester of Spanish course at this university range from 90 to 105, depending on the level.

3.4.3 Language Aptitude Test

For this study we implemented the Second Language Pragmatics Aptitude Test (SLPAT) developed by Sedaghatgoftar et al. (2019). SLPAT is composed of three parts: memory for pragmatic rule learning (20 items), mind-reading from films (10 items) and mind-reading from voices (10 items) and it also provides an overall total score. In this study all scores are transformed to 0-100 for easier presentation.

The first section of the test named "memory for pragmatic rule learning" involves measuring the ability to remember pragmatic rules from other languages which are culturally different from English. To develop the items in this section, some pioneering studies, such as Gass and Neu (1996), Wierzbicka (1985), Trosborg (2010) and Han (1992) were consulted and eight striking cross-cultural pragmatic differences between English and some other languages (Korean, Greek, Arabic, Hebrew and Japanese) were singled out. This section comprises two phases: an exposure phase - in which participants read information on cross-cultural verbal behavior of people with different nationalities - and a test phase - in which participants are tested on their ability to recognize examples of such behavior.

The second section of the test entitled "mind-reading from films" is modelled on various studies in the area of mind-reading (e.g., Baron-Cohen & Cross, 1992; Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001; Golan, Baron-Cohen, Hill, & Golan, 2006; Golan, Baron-Cohen, Hill, & Rutherford, 2006). It measures the ability to recognize the emotions and mental states in others, using social scenes from films. This section of the test consists of ten items, each displaying a fragment from a movie containing a scene considered as proper for the purposes of this study.

The third section of the test is modelled on and named after "mind-reading from voices" studies (e.g., Golan, Baron-Cohen, Hill, & Rutherford, 2006). The spectrum of emotions used in Golan, Baron-Cohen, Hill, & Rutherford (2006) was considered as the basis. Ten of the emotions were selected, the proper utterances to convey the feelings were decided upon, verbalized and recorded. The recordings are unisex (feminine) and in Persian (the native language in Iran). In order to make sure that the test-takers are not familiar with Persian, the demographic part of the test requires the participants to specify the languages they are familiar with, if any, in addition to English as their mother tongue. The purpose in doing so is to identify the participants who are familiar with Persian (or the other languages presented in the first section) and to exclude them from the study. The subjects have to listen to the recording in each test item and choose from among four choices how they think the speaker feels.

3.4.4 Motivation Scale

We adopted a motivation scale approach based on the L2 motivational self-system (Dörnyei, 2005, 2009) which stems largely from the concepts of possible selves and self-discrepancy theory. The model proposes that language learners are guided by visions of 'second language selves', one which attracts them toward becoming an idealized L2 user (ideal L2 self) and one which pushes them to learn the target language based on societal obligation or a fear of failure (ought-to L2 self).

We adopted a 33 question/6 factor version of the L2 Motivational Self System created by Kong et al. (2018). They offer the following descriptions of the motivation scale elements:

1. Ideal L2 self: "The ideal L2 self refers to a positive future image of the L2 self. For example, learners who have developed a vivid ideal L2 self are likely to endeavor to learn an L2 by imagining themselves communicating fluently using the L2 in the future."

2. Ought-to L2 self: "(This element) pushes people from societal obligation or a fear of failure."

3. International posture: "It captures a tendency to relate oneself to the international community rather than any specific L2 group. The key characteristics of international posture are described as an interest in global issues or international affairs, a willingness to travel, stay, or work abroad, and a readiness to interact with foreigners or foreign cultures."

[4] https://emmersion.ai/products/webcape/, section Efficiency

4. Competitiveness: "Competitiveness can be described as the desire to excel in comparison to others and contends that a learner constantly compares oneself with one's idealized self-image or with other learners, feels pressured to out-do other students."

5. L2 learning Experience or Attitudes: "L2 learning experience is related to the learners' environment including teachers, peer groups, curriculum, and their attitudes toward L2 learning."

6. Learners' Intended Effort or Motivated Behavior in L2 Learning: "This motivation element evaluates how much effort users are determined to make and how hard they are ready to study."

3.4.5 Global Language Score

We asked participants to complete an adapted version of the Bilingual Language Profile (Birdsong et al., 2012). It provides a Global Language Score (GLS) for any second languages spoken by the participants. The GLS is calculated on separate modules on language history, language use, language proficiency and language attitudes. GLS can vary from 0 to 218 but for this study we re-scaled it to a scale from 0 to 100.

For example, a GLS score of 218 (or rescaled as 100 percent) for English would be appropriate for participants born into an English-speaking family, in an English-speaking country, who started studying English immediately, for whom all classes at school were in English, who speak only English all the time with family, friends, and at work. Their language history and language use are entirely English- based. They feel totally proficient in English, and they identify themselves with an English-speaking culture.

3.5 Efficacy Computation

The test results (WebCAPE and TNT scores) alone cannot give a clear picture of the efficacy of a language learning app because they do not account for the time spent studying. We are therefore relying on a direct and objective measure of efficacy, which is defined as follows:

Efficacy=Improvement per one hour of study.

$$\text{Efficacy}=\frac{\text{Effect}}{\text{Effort}}=\frac{\text{Improvement of language skills}}{\text{Study time}}=\frac{\text{Final-Initial Test score}}{\text{Server Study Time}}$$

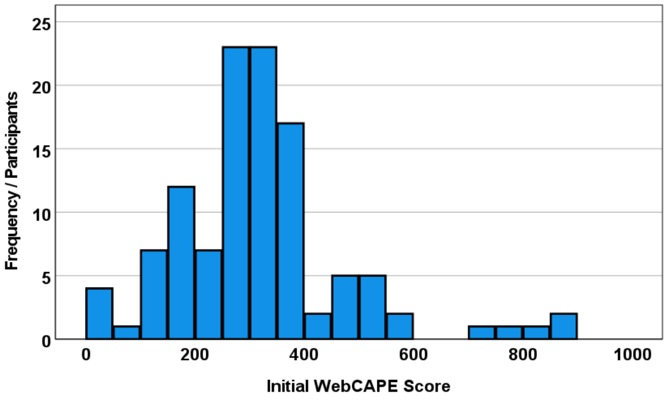
The efficacy measure includes both the amount of progress made and the amount of effort invested by each study participant. This is a direct and objective measure of efficacy: direct because it directly includes the effect and the effort; objective, because the effect is measured by an independent language test and the effort is measured by the time recorded on the computer servers.

4. Results

4.1 Reading/Grammar Proficiency

At the beginning of the study, participants had an average WebCAPE score (mean) of 316.7 (SD=161.3). This indicates that most participants in the study had already achieved a level of Spanish beyond beginner level and beyond the first college semester level as measured by WebCAPE. One of the 114 users was missing one WebCAPE measure so the WebCAPE results are based on 113 participants.

Figure 1. Initial Distribution of WebCAPE score



The initial distribution by semester level placement by WebCAPE indicated that 34.5% (n=39) were initially placed in first semester Spanish, 32.7% (n=37) in second semester and the last 32.7% (n=37) in third and fourth college semester of Spanish. In previous efficacy studies (Vesselinov, Grego et al. 2009-2020) more than 85% of the participants were initially placed in the first college semester of Spanish. The first semester level includes true beginners or novice users with a WebCAPE score between 0 and 269. The WC_CEFR initial estimate placed only 30.1% of the users at CEFR A1 Beginner level. This Busuu 2021 study sample can be characterized as intermediate level users since two thirds of the sample are above the first college semester of Spanish.

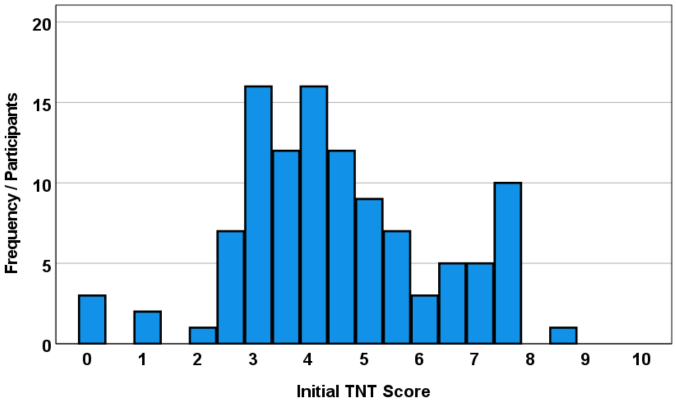
Almost 80% (79.6%) of the users improved their reading/grammar score at the end of the study compared to their initial score (95%CI 71.2-86.1). This measure was calculated as the difference between the final and initial WebCAPE scores. The average improvement was 83.7 points (95%CI 56.8-110.7). Based on the WC_CEFR difference, 56.6% of the users improved their CEFR score[5] (95%CI 47-65). More than half of the users (53.1%) increased their college semester level placement by at least one semester (95%CI 44-62). Specifically, 40.7% (n=46) moved up one college semester, 10.6% (n=12) moved up two semesters, and 1.8% (n=2) moved up three semesters.

4.2 Oral Proficiency

Five of the 114 users had less than 8 hours of study and were excluded from the analysis. The sample for the oral proficiency analysis was n=109.

The initial level of oral proficiency was on average 4.5 TNT points (out of 10) with 1.8 points standard deviation. TNT_CEFR and TNT_ACTFL initial results placed about 4% of the users at the A1 or Novice-Low level. In previous efficacy studies (Vesselinov, Grego at al., 2009-2020) more than half of the users were placed initially at ACTFL Novice Low level.

Figure 2. Initial TNT Score



About 71% (70.6%) of the users improved their TNT oral proficiency level (95%CI 62-78). The average improvement was 0.6 TNT points with 1.2 TNT points standard deviation. About 45.9% of the users improved their TNT_CEFR level (95%CI 37-55) and 47.7% improved their TNT_ACTFL level (95%CI 39-57).

4.3 Reading/Grammar and Oral Proficiency Gains Combined

In order to improve the understanding of the overall performance we combined the reading/grammar gain and oral proficiency gain. The gain is measured as difference between the final test score and the initial test score and if the difference is greater than zero, there is an improvement in this area, otherwise there is no gain. Overall, 55.6% (n=60) of the participants improved both their reading/grammar and oral proficiency (95%CI, 46-65). This is the best performing group of users in this study. About 24% (n=26) of the participants improved only their reading/grammar proficiency and 14% (n=16) improved only their oral proficiency. Only 5.6% (n=6) of the users did not improve at all. This also means that 94.4% of all participants improved in at least in one area of proficiency (95%CI 88.1% - 97.6%).

4.4 Efficacy

The reading/grammar efficacy was computed by dividing the gain in WebCAPE by the total study time. The study time includes work with the Busuu app and Live lessons. The average reading/grammar efficacy was 5.8 (95%CI 3.4-8.1). This means that on average for one hour of study the participants gained 5.8 WebCAPE points. The initial WebCAPE level was above 269 points (Semester 2). This means that on average 77 more points are needed to move from Semester 2 to Semester 3 (at least 346 points). With an average efficacy of 5.8 points per hour, a person can progress from Semester 2 to Semester 3 with 13.3 hours of study (77 divided by 5.8). In other words, the transformed reading/grammar efficacy implies that a Busuu second semester user will need on average 13.3 hours to reach the Semester 3 level with transformed 95%CI[6] : between 9.5 and 22.6 hours of study.

The oral proficiency efficacy was computed by dividing the gain in TNT by the total study time. The average oral proficiency efficacy was 0.036 (95%CI 0.02-0.05). This means that on average for one hour of study the participants gained 0.036 TNT points. The initial TNT level was 4.5 points (out of 10). With an average efficacy of 0.036 points per hour, a person can progress with one full TNT point with 27.8 hours of study (1 divided by 0.036). In other words, the transformed oral proficiency efficacy implies that a Busuu intermediate user will need on average 27.8 hours to increase their TNT level by one full point with transformed 95%CI[7] : between 20 and 50 hours of study.

[5] Note that TNT estimates for CEFR and ACTFL provide in-between levels, like A1-A2, A2-B2, ..., for CEFR, and Novice Low – Novice Mid, Novice Mid-Novice High, …, forACTFL.

[6] Lower End 77/8.1=9.5 and Upper End 77/3.4=22.6

[7] Lower End 1/0.05=20 and Upper End 1/0.02=50

4.5 Beginner and Intermediate Users Comparison

The semester placement by WebCAPE allows to classify the participants by their initial level as “beginner/novice” users (WebCAPE Semester 1) and more advanced or “intermediate” users.

Overall, the beginner users had bigger improvement in their reading/grammar and oral proficiency than the intermediate users.

The beginner users’ reading/grammar efficacy on average was 9.5 WebCAPE points per one hour of study (95%CI 4.9-14.2) compared to mean of 3.8 (95%CI 1.2-6.4) for the intermediate users (p=0.02).

On average, the beginner users gained 123 WebCAPE points (95%CI 91-155) compared to mean of 63 (95%CI 26-100) for the intermediate users (p=0.036).

Overall, 92.3% of the beginner users improved their reading/grammar proficiency (95%CI 79-98) compared to 73% for the intermediate users (95%CI 62-82), p=0.015.

About 77% of the beginner users moved up at least 1 semester in their college placement (95%CI 61-88), compared to 41% of the intermediate users (95%CI 30-52).

The differences between the two groups for the oral proficiency and gain were not statistically significant.

Overall, 75% of the beginner users improved their oral proficiency (95%CI 59-86), compared to 68.1% for the intermediate users (95%CI 57-78), p=0.4.

Finally, we combined the reading/grammar and oral proficiency gain (Yes/No) and created 4 groups of users where the participants had: 1. No improvement at all, 2. Improvement in written proficiency only, 3. Improvement in oral proficiency only, and 4. Improvement in both reading/grammar and oral proficiency (see Table 1).

All of the beginner users (100%) improved their proficiency either in reading/grammar, or oral proficiency, compared to 81.7% of the intermediate users.

About 67% of the beginner users improved in both reading/grammar and oral proficiency (95%CI 50-80), compared to 50% of the intermediate users (95%CI 39-61).

Table 1. Improvement for Beginner and Intermediate Users

	Proficiency Gain	Beginner % (n)	Intermediate % (n)	Total % (n)
1.	No improvement	0 (0)	8.3 (6)	5.6 (6)
2.	Reading/Grammar Improvement Only	25.0 (9)	23.6 (17)	24.1 (26)
3.	Oral Improvement Only	8.3 (3)	18.1 (13)	14.8 (16)
4.	Both Reading/Grammar and Oral Improvement	66.7 (24)	50.0 (36)	55.6 (60)
	Total	100 (36)	100 (72)	100 (108)

4.6 Univariate Factors Affecting the Language Proficiency Gain

The efficacy measures have two parts: the language proficiency gain in the numerator and the study time in the denominator. Any change in the efficacy can be due to changes either in the numerator or the denominator, or both. That is why for the factors' effects we are focusing on the effects of the numerator in order to clarify the analysis. The denominator (study time) is presented as a separate factor.

4.6.1 Study Time and Live Lessons

Study time is part of the efficacy computation, so its effect is already included. But for the individual progress/gain it is still valuable to explore the time effect as a factor.

- Effect on Reading/Grammar Gain

The total study time (Busuu app time plus Live lessons time) varied from 3 hours to 50 hours. The average median time was 16.5 hours (Q1=13.5 and Q3=24.3). Using the CART models, we determined that for the reading/grammar gain the optimal threshold was between15 and 22.3 hours of study. The average reading/grammar gain was 126.7 points (SD=166.3) compared to a gain of 69.5 for less than 15 hours and 48.5 for more than 22.3 hours. The effect size direction is clear, and it is statistically significant (p=0.047).

All participants were given access to 2 Live lessons a week with a professional language teacher. The study continued for 8 weeks so a total of 16 Live lessons were available for them. It is remarkable that 3 people did not book any lessons at all and 8 people paid for additional Live lessons. The number of Live lessons taken varied from 0 to 19 with the median number of lessons being 13.5 (Q1=9 and Q3=16). After the end of the study, we surveyed the participants who did not use all available Live lessons. 41 people responded and there were two main reasons for not using all the lessons. First, about 60% (n=24) of them stated that they did not have enough time for more lessons. Second, about 12% did not feel ready or prepared enough for Live lessons. The rest of the participants had technical difficulties with their computer, audio or video equipment and could not book more Live lessons. Only 2 people did not like the Live lessons enough or wanted to use their smartphone for the Live lessons which was not possible.

CART models determined that the optimal number of lessons was between 7 and 12. Participants from this group gained on average 121.7 WebCAPE points compared to 56.3 with less than 7 lessons and 68.4 with more than 12 lessons. The direction of the effect is clear, but the effect was not statistically significant (p=0.2).

It is also interesting to investigate the structure of study time, specifically the percent of total time was used for Live lessons. There was a wide variety of user preferences. The proportion of study time through Live lessons in the total study time varied from 0% to 87.7%. The highest percentage was a user who used the Busuu app for 1 hour but liked the Live lessons so much that they booked 15 Live lessons. The median was about 32% (Q1=22.6% and Q3=40%). CART models determined that the optimal ratio of lessons was between 30.4% and 34.6%. Participants from this group gained on average 159.1 WebCAPE points compared to 62.3 with less than 30.4% and 78.9 with more than 34.6%. The direction of the effect is clear, and the effect is marginally statistically significant (p=0.09).

- Effect on Oral Proficiency Gain

The CART model confirmed similar optimal points for the oral proficiency gain regarding total study time. Users with 15 to 22.3 hours of study had the highest gain on TNT with an average of 0.89 TNT points (SD=1.3) compared to 0.38 points for less than 16 hours and 0.46 points for more than 22.3 hours.

Live lessons are expected to have a big impact on the oral proficiency gain. The number of lessons is linearly related to the oral proficiency gain but with threshold points. Users with 13 or more lessons had an average gain of 0.5 TNT points, followed by users with 3 to 12 lessons with gain of 0.36 TNT points. Users with less than 3 lessons had the lowest gain of 0.33 TNT points. The differences were not statistically significant, but the trend is present.

CART models determined that the optimal proportion of Live lessons for oral proficiency gain was between 31.7% and 48.6%. Participants from this group gained on average 0.67 TNT points compared to 0.61 points with less than 31.7% and 0.2 points gain for more than 48.6% (n=10). Basically, study time with more than 48.5% Live lessons is less beneficial but the differences are not statistically significant (p=0.5) because the last group is very small.

4.6.2 Initial Level and Language Proficiency

From previous efficacy studies (Vesselinov, Grego et al., 2009-2020) we know that the initial level of language knowledge has a clear effect on proficiency gain. True beginners or novice users gain faster than users with higher initial level of proficiency. The new factor in this study is the fact that this is the first study with users predominantly at an intermediate level. Most of the previous studies were based on novice users.

As expected, the initial level of reading/grammar proficiency is inversely related (Figure 3) to the gain in Reading/Grammar (r = -0.45).

Figure 3. Effect of the Initial WebCAPE Score on Reading/Grammar Gain

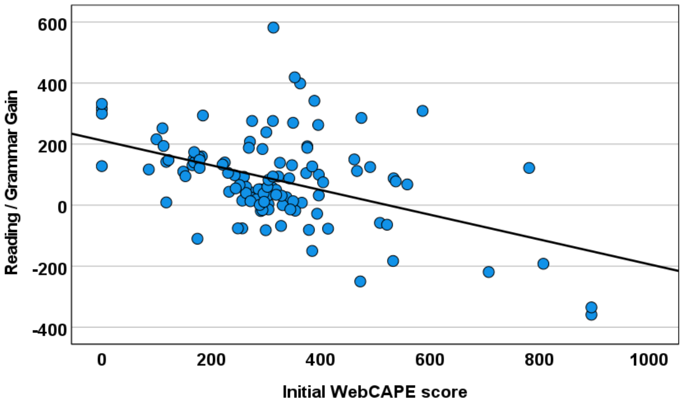
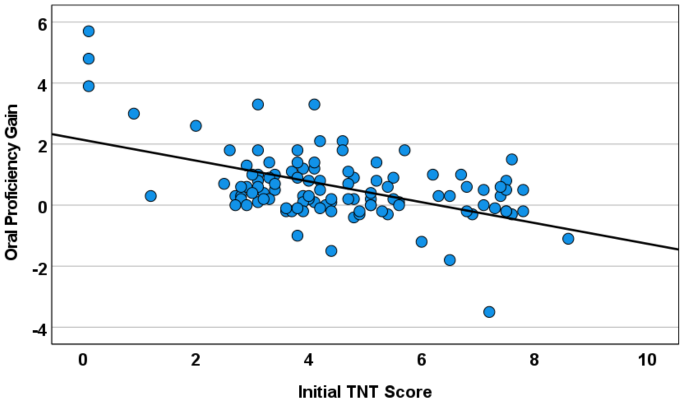


Figure 4. Effect of the Initial TNT Score on Oral Proficiency Gain



The initial level of oral proficiency is inversely related (Figure 4) to the gain in oral proficiency (r = -0.5). As expected, the gains in reading/grammar and oral proficiency are greater for users with lower initial language proficiency level.

4.6.3 Motivation

The level of motivation of the participants in this study was very high. The median level of the total motivation was at 74% out of a maximum of 100% (Q1=68% & Q3=79%). Five of the six components of the total motivation were also very high. The median for motivation elements "Ideal Self", "Learning Attitude" and "Intended Effort" was 80%. The median for motivation elements "International Posture" and "Competitiveness" was 77%. Only motivation element "Ought to Self" had a lower median level at 46% (Q1=34%, Q3=60%) which is in line with the results from previous studies (Vesselinov, Grego et al., 2019). This level suggests that the participants were not motivated by fear of failure and they were not that susceptible to pressure from societal obligation.

The effect of motivation is not linear in nature and has some concentration thresholds and inflection points. This means that being more motivated to a certain extend does not necessarily lead to better gain in language proficiency. This fact has been discovered in other efficacy studies as well (Vesselinov, Grego et al., 2019).

We used CART models to investigate this effect. The total motivation effect on Reading/Grammar Gain had a threshold of 88.2%. Users at this or higher motivation levels had on average 192.3 WebCAPE points gain compared to a gain of 75.4 points for lower motivation and this difference was statistically significant (p=0.028).

The threshold for total motivation effect on Oral Proficiency Gain was at 90.5% or higher. Users at this motivation level had on average 1.2 TNT points gain, compared to 0.58 points for lower motivation but the difference was not statistically significant (p=0.27).

4.6.4 Language Aptitude

The SLPAT aptitude test (Sedaghatgoftar et al., 2019) was available only in English so the sample size for this analysis was n=85. This test was not offered to the Portuguese participants. The overall aptitude median level was at 60% (Q1=50 & Q3=70%). The three aptitude elements were as follow: "Rule" at 60% median level, "Movie" at 50% and "Voice" at 70%.

The overall aptitude score has an effect on both reading/grammar gain and the oral proficiency gain. We found a concentration threshold level of the effect on reading/grammar gain at 71.3%. Users with an aptitude level of 71.3% or higher had on average gain of 124.5 WebCAPE points compared to 69.3 points for users with lower aptitude level (p=0.046).

We also discovered a concentration threshold point at 73.8% total aptitude level for the oral proficiency gain. Users with 73.8% and higher aptitude had on average a gain of 1.4 TNT points compared to 0.53 points for users with a lower aptitude level (p=0.01).

4.6.5 Second Language Profile

We asked participants who said that they know a second/foreign language to complete the GLS. Overall, 39 people successfully completed the GLS and this was the basis for our analysis. From previous studies (Vesselinov, Grego, et al., 2019) we know that GLS for the participants' native language is very close to 100% with little variation and is not a very interesting factor. But the GLS for the second language is a fair potential factor.

Overall, the total GLS score had a median value of about 33% (Q1=25% & Q3=48%). The values for GLS varied from 2% to 82%. GLS does not have a linear relationship with either of the language gains (WebCAPE or TNT). CART discovered two threshold values for GLS. Users with a GLS score of 30.2% or higher had a reading/grammar gain of 92.3 WebCAPE points compared to 50.9 points for people with a lower GLS score. The threshold for oral proficiency gain was at 49.3%. Users with this GLS level or higher had an average oral proficiency gain of 1.2 TNT points compared to 0.48 points for users with a lower GLS. Both sets of differences were not statistically significant, but the direction of the effect is present. Higher GLS score, i.e., better knowledge of a second language improves the language proficiency gains but the small sample (n=39) does not have enough statistical power.

4.6.6 Crosslinguistic influence

The Portuguese part of our sample (n=28) performed better than the English part (n=86) in reading/grammar gain. They had an average gain of 92.9 WebCAPE points compared to 80.9 points, but the difference was not statistically significant (p=0.7). In oral proficiency gain the Portuguese participants performed worse than the English participants. They had an average oral proficiency gain of 0.42 TNT point compared to 0.66 points and this difference was not significant (p=0.36).

In the English sample alone, there were 10 non-native speakers. Interestingly enough, the results were the opposite of the English vs Portuguese comparison. The English native speakers performed better than the non-native speakers of English in reading/grammar gain but were worse in oral proficiency gain. Neither of the differences was statistically significant.

4.6.7 Demographics and Other Factors

Based on the CART models, we could divide age into 3 groups (18-25), (26-40), and (>40). For reading/grammar gain, the results for the middle group (26-40) were slightly better than the other two but no significant differences were found. For oral proficiency gain, the results were best for the youngest group 18-25), followed in order by the other two with no significant differences.

The female participants performed slightly better than male on reading/grammar gain but were worse on oral proficiency gain with no statistical significance.

Participants who studied Spanish for business/work performed better on reading/grammar gain but were worse on oral proficiency gain. Participants who studied for personal interest showed the best oral proficiency gain.

Participants with higher levels of education tend to get higher oral proficiency gains. The effect on reading/grammar gain was not clear.

Part-time and unemployed participants had the highest results in oral proficiency gain. For reading/grammar gain the pattern is not that clear.

The language environment contributed to some extent to the study process. Users who had a close friend or spouse who spoke Spanish tend to have better oral proficiency gain, but the difference was not statistically significant. Parents and grandparents who spoke Spanish were too few (n=7) for measurable effect.

Participants who claimed to know a second language did not perform better. Only users who had a good knowledge of the second language (higher GLS) tend to get better results. Just knowing a little bit of another language is not a strong enough incremental factor for improvement.

Growing up in a multilingual family has a strong effect on oral proficiency gain. Users from such families had an average gain of 1.4 TNT points compared to 0.5 points for the other group (p=0.035). Living for a longer period of time (6 months or more) in a foreign language country has some positive effects on the oral proficiency gain but it is not statistically significant.

At the end of the study the participants were asked about their satisfaction with the Busuu app and the Live lessons. These are potential indicators or factors for their performance as well. The questions asked were about how easy, helpful, enjoyable, and satisfactory their experience was. The answers were on a Likert scale (1=strongly disagree to 5 strongly agree), which later was recodes as “Yes” (5=strongly agree and 4=agree) and “No” (the rest 1-3). The Busuu app was liked universally with answers of “Yes” between 92.5% and 95.3%. Of course, this is expected since the participants were randomly selected from existing Busuu users.

Participants also thought that Live lessons were easy (92%), helpful (91%) and enjoyable (91%) but were a little less satisfactory (87%). Users who were more satisfied with the lessons had slightly better results on both reading/grammar and oral proficiency gains, but the differences were not statistically significant.

4.7 Multivariate Models of Success

The univariate factors' effects carry important but limited information. They show the effect of one factor at a time with no consideration for other factors. For example, we showed the effect of motivation for language proficiency gain without taking into account other potentially important factors like the initial level of language knowledge, second language profile (GLS), etc. This situation is not very realistic. In order to evaluate the more realistic factor effects we built CART multivariate models.

The first CART model we built had a binary dependent variable (outcome): 1=Improved both reading/grammar and oral proficiency, and 0=Did not improve. Improvement or gain is defined and present (Yes) when the final test score is higher than the initial test score and "No gain" otherwise. The model had very good ROC AUC of 83.9%. The Variable Importance Measure (VIM) showed that the most important factor is Language Aptitude, followed by GLS, and initial oral proficiency (see Table 2).

Table 2. Variable Importance Measure (VIM) for Increase both Reading/Grammar and Oral Proficiencies

Rank	Factor	VIM
1	Language Aptitude	100.0
2	Second Language Profile (GLS)	71.3
3	Initial Oral Proficiency	71.1
4	Employment Status	42.2
5	Initial Reading/Grammar Proficiency	41.4
6	Age	34.0
7	Total Study Time	32.5
8	Number of Lessons	21.8
9	Study time on Busuu app	16.2
10	Percent Lessons	14.9
11	Total Motivation	11.7
12	Native Language (Portuguese)	8.3
13	Education	2.3

The second CART model we built had a binary dependent variable (outcome): 1=Improved reading/grammar proficiency, and 0=Did not improve. The model had a very good ROC AUC=87.4%. The VIM showed that the most important factor is initial reading/grammar proficiency, followed by language aptitude, and initial oral proficiency level (see Table A3 in the Appendix).

The third CART model we built had a binary dependent variable (outcome): 1=Improved oral proficiency, and 0=Did not improve. The model had a very good ROC AUC of 90.5%. The VIM showed that the most important factor is second language profile (GLS), followed by initial oral proficiency level and study time on the Busuu app, etc. (see Table A4 in the Appendix).

Overall, the top 3 factors for improving language proficiency are language aptitude, second language profile and initial level of language proficiency.

The first CART model also discovered some more pronounced but generalizable paths to success (see Table 3).

The other two CART models also produced paths to success for an increase in one of the proficiencies and the results are presented in Tables A5 and A6 in the Appendix.

Table 3. Path to Success: Increase both Reading/Grammar and Oral Proficiencies

Path	Success Rate (%)	Description
1	99.0	Initial TNT >3.5 and employment: retired or other type of employment
2	90.9	Initial TNT ≤3.5 and Total Study Time ≤ 22.7 hours
3	80.0	Initial TNT >3.5 and GLS >47.2 and Percent Lessons > 40.6 and employment: unemployed, full time or part time employed, or homemaker
4	72.7	Initial TNT >3.5 and GLS (23.8-47.2] and Language Aptitude >68.8% and employment: unemployed, full time or part time employed, or homemaker
5	63.6	Initial TNT >3.5 and GLS ≤ 23.8 and Initial WebCAPE > 242 and employment: unemployed, full time or part time employed, or homemaker

5. Discussion and Conclusion

This is our first efficacy study with predominantly intermediate language learners of Spanish. Overall, about 80% of all participants improved their reading/grammar proficiency and about 71% improved their oral proficiency. Combining the two proficiency gains showed that 94.4% of the participants improved in at least one of them. More than half of the participants (55.6%) improved in both oral and reading/grammar proficiency.

The reading/grammar efficacy of Busuu for intermediate users was measured as a gain of 5.8 WebCAPE points per one hour of study. Based on the WebCAPE requirements for third semester Spanish, we can say that it would take Busuu users on average about 13 hours of study to advance from second to third semester Spanish (95%CI 10-23). About 53% of the users increased their college placement semester at least by one semester.

The Busuu oral proficiency efficacy was a gain of 0.036 TNT points per one hour of study. In other words, an intermediate Busuu users on average would need 27.8 hours (95%CI 20-50) to increase their oral proficiency by one full level (out of 10 TNT levels).

The median total study time was 16.5 hours which is in line with our previous studies (Vesselinov, Grego, et al., 2009-2020). These studies showed that language app users on average spend about 1-2 hours a week studying foreign language. There are some studies (e.g. Jiang et al., 2020) that base their evaluation on study hours in the range of 100-120 hours. Given the rate of 2 study hours per week, this would require about one year of study.

In this study we first evaluated the univariate effect of different factors. The most interesting finding is that most effects have a threshold that triggers the effect, i.e., the effects are not linear in nature. The total study time for example, is most effective in the range of 2 to 3 hours a week for both reading/grammar and oral proficiency gain. Less or more study time still improves the results but at a lower rate. The optimal range of the Live lessons turns out to be from 1 to 1.5 lessons a week for the reading/grammar proficiency gain. For oral proficiency gain, the more Live lessons the better with a suggested lower limit of 1.6 Live lessons per week. The structure of the study time also matters. The optimal portion for Live lessons is roughly between 30% and 50% of the total study time.

The initial level of language proficiency is one of the most important factors for proficiency gains. Less advanced users tend to improve their proficiency at a higher rate for both reading/grammar and oral proficiency.

Motivation of the participants is a factor for their success, but it has a steep threshold level. The effect of motivation is more pronounced at about the 90% level. Users with about 90% motivation and above get better results than the rest, other things being equal.

Language aptitude is one of the strongest factors for success. The higher the aptitude score, the better. Users with more than 70% language aptitude tend to have better results, other things being equal.

Simply speaking a second language is not a very strong factor but knowing the second language well is a strong factor. Users with second language profile (GLS) of 30% and more did better on reading/grammar proficiency gain. Users with 50% or better did better on oral proficiency gain. Demographics factors, education, employment, etc. did not have a clear directional effect of the language proficiency gain.

Finally, based on the multivariate CART models we determined the most influential factors. For simultaneous increase in both reading/grammar and oral proficiency the most important factors are language aptitude, second language profile (GLS), initial oral proficiency level plus all study time related measures, including total time, number of lessons, time with Busuu app, and percent Live lessons. We also found paths to success rates with user groups demonstrating high success rate from 64% to 99%. There are different paths for users to improve their language proficiency. For example, some may have lower initial language proficiency level so they may only need to increase their study time. Other users with higher initial language proficiency can be helped with a good knowledge of a second language (higher GLS), higher language aptitude, or improved structure of their study time (e.g., more Live lessons). Some of the factors are not easily changeable like language aptitude and GLS, but motivation and all study time elements can be changed, particularly the structure of the study time.

Limitations of the Study

This study was funded by Busuu and participants were existing users of Busuu who were incentivized to take part in the study through the provision of free language lessons. Both of these limitations may have led to a degree of bias. For example, this cohort of learners may have been more motivated than a typical group. However, there are currently very few opportunities for language app developers to benefit from independent research studies into their efficacy, and so the current best way for them to give confidence to their customers is through commissioning research with engaged groups of users.

One of the standardized tests (WebCAPE) used in this study was administered in English, which may have affected the ability of the Portuguese-speaking participants to understand exactly what to do in order to complete the test activities and may have impacted their results accordingly.

The study also lacked a control group or comparison group to contrast results with.

Recommendations for Future Research

Previous efficacy studies into language learning apps have tended to focus on Spanish. Future studies would benefit from investigating the efficacy of additional languages, including English.

For greater confidence in the results, future studies could attempt to utilize a comparison study which contrasts results from an app such as Busuu with results from a traditional in-person college semester of language study.

Acknowledgements

We want to thank Kirsten Campbell-Howes, M.S. and Ziwei “Jo” Huang, M.S. for their contribution to the Busuu teaching approach description and the literature review.

Special thanks are due to Juliana Matos for technical support of the study.

We would like to thank Busuu for providing funding for this study and securing the study time records from their servers.

References

Alonso, R. A. (2019). Crosslinguistic Influence in Second Language Acquisition. The Encyclopedia of Applied Linguistics, 1–7. doi:10.1002/9781405198431.wbeal0292.pub2

Baron-Cohen, S., & Cross, P. (1992). Reading the eyes: Evidence for the role of perception in the development of a theory of mind. Mind and Language, 7 (1&2), 172–186.

Baron-Cohen, S., Golan, O., Wheelwright, S., & Hill, J. J. (2004). Mind Reading: The Interactive Guide to Emotions. London: Jessica Kingsley Limited.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The ''Reading the Mind in Films'' task: Complex emotion recognition in adults with and without autism spectrum conditions. Journal of Child Psychology and Psychiatry, 42 (2), 241–251.

Birdsong, D., Gertken, L., & Amengual, M. Bilingual Language Profile: An Easy-to-Use Instrument to Assess Bilingualism. COERLL, University of Texas at Austin. Web. 20 Jan. 2012. https://sites.la.utexas.edu/bilingual/.

Breiman, L., Friedman, J., Stone, C., Olsen, R., Classification and Regression Trees, Chapman & Hall/CRC, 1984.

Burston, J. (2014). MALL: The pedagogical challenges. Computer Assisted Language Learning, (27)4, 344-357. https://doi.org/10.1080/09588221.2014.914539

Burston, J. (2015). Twenty years of MALL project implementation: A meta-analysis of learning outcomes. 27(1), 4-20. https://doi.org/10.1017/S0958344014000159

BYU 2021 Academic Calendar. (2021). Retrieved January 22, 2021, from https://enrollment2.byu.edu/academic-calendar BYU Class Search. (2021). Retrieved January 22, 2021, from http://saasta.byu.edu/noauth/classSchedule/index. .php?yearTerm=20213&amp;curriculumId=05294&amp;titleCode=019

Carroll, J. B., & Sapon, S. M. (1959). Modern language aptitude test.

Cawalho, A. M., & Silva, A. J. B. (2006). Cross-Linguistic Influence in Third Language Acquisition: The Case of Spanish-English Bilinguals' Acquisition of Portuguese. Foreign Language Annals, 39(2), 185–202. doi:10.1111/j.1944-9720.2006.tb02261.x

Chinnery, G. (2006). Going to the MALL: Mobile Assisted Language Learning. Language Learning & Technology, 10(1), 9-16. https://scholarspace.manoa.hawaii.edu/bitstream/10125/44040/1/10_01_emerging.pdf

Dörnyei, Z. (2010). The relationship between language aptitude and language learning motivation: Individual differences from a dynamic systems perspective. In E. Macaro (Ed.), Bloomsbury Companion to Second Language Acquisition (pp. 247-267). London: Bloomsbury.

Dörnyei, Z., 2005. The psychology of the language learners. Mahwah, NJ: Lawrence Erlbaum.

Dörnyei, Z., 2009. The L2 motivational self-system. In Z. Dörnyei, & E. Ushioda (Eds.), Motivation, language identity and the L2 self (pp. 9e42). Bristol, UK: Multilingual Matters.

Gass, S. M., & Neu, J. (Eds.). (1996). Speech Acts across Cultures: Challenges to Communication in a Second Language. New York: Mouton de Gruyter.

Golan, O., Baron-Cohen, S., Hill, J. J., & Golan, Y. (2006). The ''Reading the Mind in Films'' task: Complex emotion recognition in adults with and without autism spectrum conditions. Social Neuroscience, 1 (2), 111–123.

Golan, O., Baron-Cohen, S., Hill, J. J., & Rutherford, M. D. (2006). The 'Reading the Mind in the Voice' test revised: A study of complex emotion recognition in adults with and without autism spectrum conditions. Journal of Autism and Developmental Disorders, 37 (6), 1096–1106.

Habing, B., Grego, J., Vesselinov, R., Predictive and Psychometric Properties of the TrueNorth Test (TNT), 2020, http://comparelanguageapps.com/documentation/TNT_Final_Report.pdf .

Han, C. (1992). A comparative study of compliment responses: Korean females in Korean interactions and in English interactions. Working Papers in Educational Linguistics, 8 (2), 17–31.

Hastie, T., Tibshirani, R., Friedman, J., The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics), 2009.

Huang, Z. (2020) Thirteen years since the first iPhone: A systematic review on the effectiveness of language learning apps on smart devices [Unpublished Master's thesis]. University of Oxford. Retrieved from https://ora.ox.ac.uk/objects/uuid:7b58715e-9f64-48a1-a28b-d2076065fefe

Jiang, X., Rollinson, J., Plonsky, L., Pajak, B., Duolingo Efficacy Study: Beginning level courses equivalent to four university semesters, https://duolingo-papers.s3.amazonaws.com/reports/duolingo-efficacy-whitepaper.pdf.

Kong, J., Han, J., Kim, S., Park, H., Kim, Y., Park, Hy. L2 Motivational Self System, international posture and competitiveness of Korean CTL and LCTL college learners: A structural equation modeling approach, System, Volume 72, February 2018, 178-189.

Kukulska-Hulme, A., & Traxler, J. (2005). Mobile learning: A handbook for educators and trainers. Routledge.

Macaro, E., Vanderplank, R., &amp; Murphy, V. A. (2010). A Compendium of Key Concepts in Second Language Acquisition. In E. Macaro (Ed.), Bloomsbury companion to second language acquisition (pp. 29-106). London: Bloomsbury.

Rosell-Aguilar, F. (2018). Autonomous language learning through a mobile application: A user evaluation of the busuu app. Computer Assisted Language Learning, 31(8), 854-881. https://doi.org/10.1080/09588221.2018.1456465

Sáfár, A., & Kormos, J. (2008). Revisiting problems with foreign language aptitude. International Review of Applied Linguistics in Language Teaching, 46(2), 113-136.

Sedaghatgoftar, N., Karimi, M., Babaii, E., Reiterer, S., Developing and validating a second language pragmatic aptitude test, Cogent Education, 2019, 6:1654650

Sparks, R., & Ganschow, L. (2001). Aptitude for learning a foreign language. Annual Review of Applied Linguistics, 21. doi:10.1017/s026719050100006x

Trosborg, A. (Ed.). (2010). Pragmatics across Languages and Cultures. New York: Mouton de Gruyter.

Vesselinov, R., & Grego, J. (2016). The Busuu efficacy study. Retrieved December 1, 2020, from http://comparelanguageapps.com/documentation/The_busuu_Study2016.pdf

Vesselinov, R., Grego, J., et al., 2009-2020, Language Efficacy Studies reports, http://comparelanguageapps.com/ .

Wierzbicka, A. (1985). Different cultures, different languages, different speech acts: Polish vs. English. Journal of Pragmatics, 9, 145–178.

Appendix

Figure A1. Sample Selection Tree

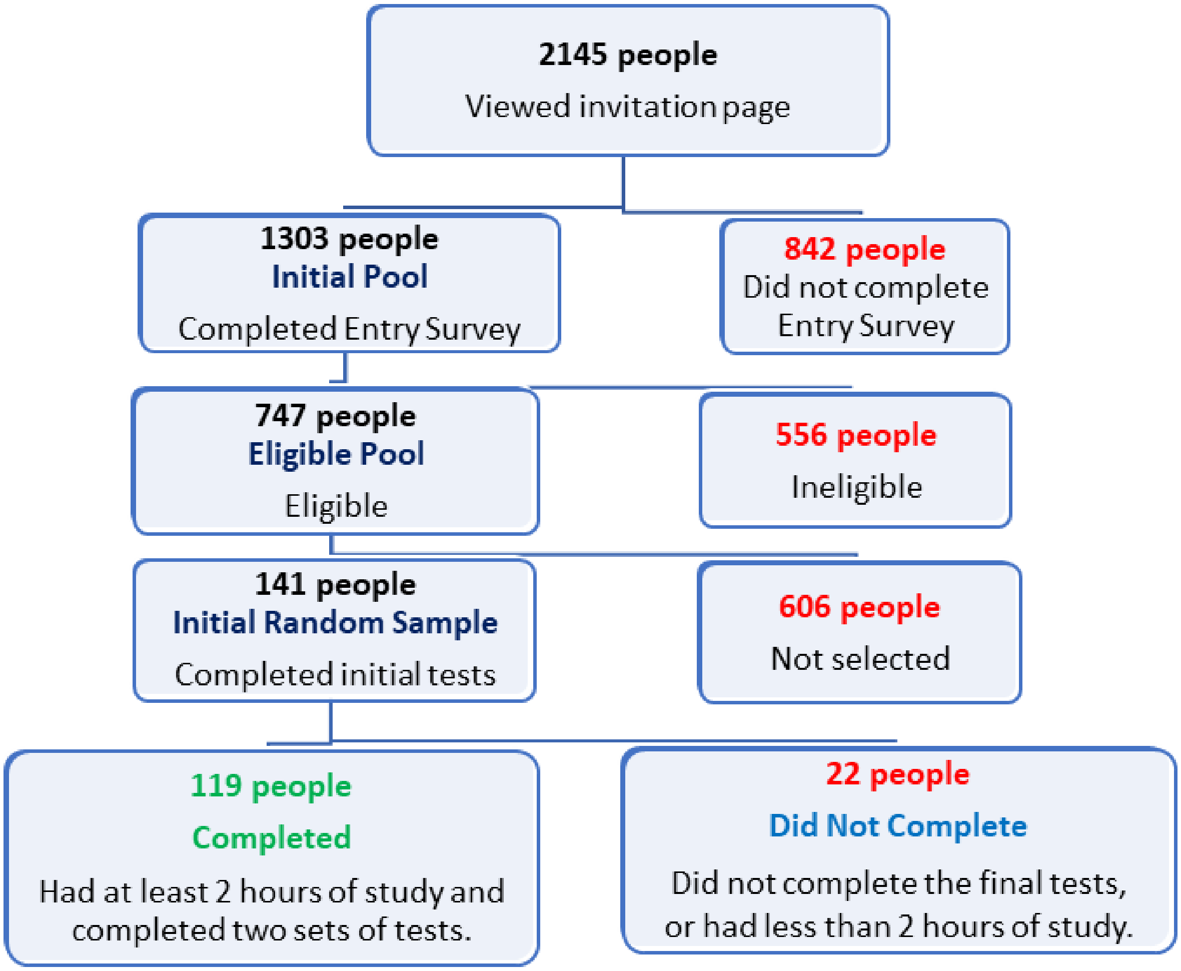


Table A1. Background Information on the Participants

Categories	Mean or %	(SD) or n	Total N
Age (mean)	38.1	(12.4)	114
Female (%)	46.8	51	109
Education (%)			113
Less than High School	1.8	2	
High School	2.7	3	
Started college but did not graduated	13.3	15	
B.A. degree	46.0	52	
Started graduate school but did not graduate	10.6	12	
M.A. degree	18.6	21	
Ph.D. degree	7.1	8	
Employment (%)			113
Unemployed	12.4	14	
Part Time	13.3	15	
Full Time	62.8	71	
Retired	3.5	4	
Homemaker	3.5	4	
Other	4.4	5	
Second Language (%)	42.1	48	114
Resident Country (%)			114
Brazil	24.6	28	
English Speaking Country	75.4	86	
Australia		2	
Canada		7	
New Zealand		1	
UK		28	
US		48	
Reason for Studying Spanish (%)			114
Business/Work	14.9	17	
Travel	23.7	27	
School	1.8	2	
Personal Interest	59.6	68	
Have close friend or spouse who speaks Spanish (%)	22.8	26	114
Have parents of grandparents who speak Spanish (%)	6.1	7	114
Spent 6 months+ in foreign language country (%)	16.7	19	114
Grew up in multilingual family (%)	9.6	11	114

Table A2. Background Scale Measures

Instrument	Median	Q1-Q3	Total N
Second Language Profile (GLS) (%)			75
Total Score	32.6	24.8-48.3	
History	6.7	2.5-33.3	
Language Use	6.0	2.0-20.0	
Proficiency	54.2	37.5-70.8	
Attitude	58.3	45.8-75.0	
Motivation (%)			113
Total	73.9	68.4-79.4	
Ideal Self	80	70-90	
Ought to Self	45.7	34.3-60.0	
International Posture	76.7	71.7-83.3	
Competitiveness	76.7	66.7-83.3	
Learning Attitude	80	80-90	
Intended Effort	80	73.3-90	
Language Aptitude (%)			85
Total	60	50-70	
Rules	60	45-82.5	
Movies	50	40-60	
Voice	70	50-80	

Table A3. Variable Importance Measure (VIM) for Increase in Reading/Grammar Proficiency

Rank	Factor	VIM
1	Initial reading/grammar proficiency	100
2	Language Aptitude	57.5
3	Initial Oral Proficiency	45.6
4	Study time on Busuu app	24.9
5	Total Study Time	20.7
6	Number of Lessons	18.7
7	Age	18.0
8	Total Motivation	16.6
9	Education	15.2
10	Percent Lessons	13.7
11	Second Language Profile (GLS)	7.9
12	Employment Status	7.3
13	Reason for Studying Spanish	4.1

Table A4. Variable Importance Measure (VIM) for Increase in Oral Proficiency

Rank	Factor	VIM
1	Second Language Profile (GLS)	100.0
2	Initial Oral Proficiency	95.8
3	Study time on the Busuu app	46.2
4	Percent Lessons	37.4
5	Total Study Time	31.4
6	Education	30.0
7	Number of Lessons	22.7
8	Gender	20.3
9	Have close friend or spouse who speaks Spanish	20.2
10	Knows second language	16.2
11	Age	15.2
12	Total Motivation	12.0
13	Reason for Studying Spanish	11.7
14	Language Aptitude	9.1
15	Initial Reading/Grammar Proficiency	4.5

Table A5. Path to Success: Increase Reading/Grammar Proficiency

Path	Success Rate (%)	Description
1	99.0	Initial WebCAPE > 290 and Initial TNT > 4.4 and Language Aptitude >71.3%
2	95.0	Initial WebCAPE (290 – 405] and Language Aptitude (48.6 – 71.3] and education: all types except unfinished college and M.A.
3	93.8	Initial WebCAPE ≤ 290

Table A6. Path to Success: Increase Oral Proficiency

Path	Success Rate (%)	Description
1	99.0	Study time with Busuu app >27.2 hours
2	99.0	Initial TNT (3.5-4.9] and Study time with the Busuu app ≤ 27.2 hours and Percent Lessons > 27.4 and GLS>29
3	93.8	Initial TNT ≤ 3.5
4	87.5	Initial TNT >3.5 and Study time with Busuu app ≤ 27.2 hours and Percent Lessons > 27.4 and GLS ≤ 29 and speaking second language
5	81.3	Initial TNT >4.9 and Study time with Busuu app ≤ 27.2 hours and Percent Lessons > 27.4 and GLS>29 and education: BA degree