The busuu Efficacy Study

FINAL REPORT

RESEARCH TEAM

ROUMEN VESSELINOV[1,2], PhD

Economics Department
Queens College,
City University of New York
roumen.vesselinov@qc.cuny.edu

JOHN GREGO, PhD

Statistics Department
University of South Carolina
grego@stat.sc.edu

May 2016

[1] Corresponding author.
[2] This report represents the individual opinion of the authors and not necessarily of the two institutions.

EXECUTIVE SUMMARY

This study was independently conducted by the Research Team from February 2016 to April 2016. A random representative sample of 196 busuu users was drawn. The participants took one college placement Spanish language test and one oral proficiency test, then studied Spanish with busuu for two months and took the same tests again. Both tests were proctored. The improvement in language abilities was measured as the difference between the final and the initial language test results. The efficacy of busuu was measured as written proficiency improvement per one hour of study and the proportion of users who improved their oral proficiency.

MAIN RESULTS

Written Proficiency Gain:

• Overall 84% of the participants improved their written proficiency.

• busuu users need on average 22.5 hours of study in a two-month period to cover the requirements for one college semester of Spanish.

Oral Proficiency Gain:

• Over 75% of busuu users increased their oral proficiency by at least one level.

SUPPLEMENTARY RESULTS

Written Proficiency:

• The efficacy of busuu is a gain of about 12 test points per one hour of study. For beginners the gain is 13.6 points and for advanced users it is 3 points.

• About 42% of the participants moved up at least one college semester level. Of those, 26% moved up one semester, 10% - two semesters, and 6% - three semesters.

Oral Proficiency:

• Over 75% of the participants increased their oral proficiency by at least one level. Of those, 44% moved up one level, 25% - two levels, and 7% - more than two levels.

User Satisfaction:

• The majority of users thought that busuu was easy to use (86%), helpful (84%), enjoyable (78%), and they were satisfied with it (74%).

• busuu received a positive Net Promoter Score of 8.4 from users.

• busuu efficacy was not affected by gender, race, age, education, native language, device used, etc.

CONTENTS

1. Introduction

There are many language learning apps available today. There are different claims from all of them about how effective or how fast one can learn a foreign language. It is very difficult to select the right app just based on their claims. Our Research Team started evaluating language learning software back in 2008. Since then we have evaluated several learning software providers and this study adds one more evaluation and tests one more claim from a language learning app.

There is a growing interest in evaluating the efficacy (or effectiveness) of language learning apps. New users, investors, analysts and academics are eager to learn what they can expect to gain by using a particular software and which software is most effective. Our research team has already conducted several studies attempting to directly evaluate the efficacy, attitude and motivation of some popular language learning software packages, namely Rosetta Stone®, Aurolog® and Berlitz®, Duolingo® and a new language app (Vesselinov 2008, Vesselinov et al. 2009a, 2009b, Vesselinov & Grego, 2012 & 2016). Since the 2012 study we regularly receive inquiries from the US and all over the world: e.g. a school district administrator in New York and in China, a foundation in India related to school excellence, major investment groups, individual users, etc. All of them want the same thing: they need help to decide which language app they should use. Other things being equal (e.g. price, appearance, ease to use, etc.) they needed independent evaluation of the efficacy of the apps and the more specific the measure is, the better.

With this study we are trying to evaluate the efficacy of a well-known language software product: busuu[3]. The company was founded in 2008 by two European entrepreneurs who believed that existing online language learning programs were missing an important social element. busuu was designed to combine self-paced language study with a supportive social network of language learners around the world. Students learn vocabulary and grammar in

[3] www.busuu.com

thematic lessons, and then put the language they have learned into practice through writing and speaking exercises which are marked by native speaker members of the community.

As of early 2016, the busuu community numbers over 60 million (per www.busuu.com) language learners around the world; there are courses in 12 languages, delivered through web, iOS and Android apps. Up to 100,000 new learners join the site each day according to the company.

Some of busuu's key features include:

• Interactive vocabulary and grammar lessons with audio, translation and multiple practice exercises;

• Audio recordings of each vocabulary item, plus example sentences and dialogues to place vocabulary and grammar in context;

• Voice recording exercises to drill pronunciation and allow students to insert their voice into a dialogue and get feedback from native speakers;

• Translations of key vocabulary, instructions and grammar tips into thirteen languages;

• Writing exercises which receive instant corrections from native speakers in the busuu community;

• busuu-talk (web only) which allows students to find language partners and practice speaking or text chatting with them;

This study was funded by busuu but the data collection and the analysis were carried out independently by the Research Team. The two language tests used in the study were designed and developed by two external independent testing companies.

2. Research Design

The random sample for this study was selected from existing or new busuu users who lived in or around London, UK or New York, US. The geographical restrictions were imposed because the test for oral proficiency must be proctored. There were some additional requirements for the potential participants. They had to be:

- Willing to study Spanish using only busuu for two months, and come to the testing location for two sets of language tests;
- At least 18 years of age;
- Not of Hispanic origin;
- Not advanced learners of Spanish.

The last requirement was due to the fact that the written language placement test used in the study has placement in college Semester 4+ as its highest evaluation group and it has limited abilities for very advanced users. The oral proficiency test has no limitations because the evaluation is done by two independent raters.

The recommended goal for the participants in the study was to use busuu for at least 16 hours during the two-month study, or two hours per week. Based on our experience with previous studies we imposed a threshold of at least two hours of study for the written test. People with less than two hours of study were not allowed to complete the study because there was not a sufficient effort for measurable progress. For the oral proficiency test the requirement was at least 16 hours of study.

The Spanish language was selected as one of the more popular languages and also because of the existence of previous research on Spanish for other language learning apps. The length of the study was approximately 8 weeks and it was conducted between the months of February 2016 and April 2016. People who successfully completed the study were given a lifetime free subscription to the premium edition of busuu for themselves and one friend of theirs. The participants in the oral proficiency tests received an official certificate for their level. No monetary or other incentives were offered to the participants.

The main instrument for evaluating the level of knowledge of Spanish was the Web Based Computer Adaptive Placement Exam[4] (WebCAPE test). It is an established university placement test and it is offered in ESL, Spanish, French, German, Russian and Chinese. It was created by Brigham Young University and is maintained by the Perpetual Technology Group. A more detailed description of the test can be found at their website[5].

The Spanish WebCAPE test has a very high validity correlation coefficient (0.91) and very high reliability (test-retest) value of 0.81. The test is adaptive so the time for taking the test varies with an average time of 20-25 minutes. The WebCAPE test gives a score (in points) and based on that score places the students in different level groups (college semesters).

Table 1. Spanish WebCAPE Test Cut-off Points

WebCAPE Test Points	College Semester Placement
Below 270	Semester 1
270-345	Semester 2
346-428	Semester 3
Above 428	Semester 4+

The WebCAPE results alone cannot give a clear picture about the efficacy of the language learning app because they do not account for the time spent studying. That is why we are relying on a **direct and objective** measure of efficacy which is defined as follows:

$$Efficacy = \frac{\text{Effect}}{\text{Effort}} = \frac{\text{Improvement of language skills}}{\text{Study time}} = \frac{\text{Final-Initial WebCAPE test score}}{\text{Hours of study}}$$

Efficacy=Improvement per one hour of study

This measure includes both the amount of progress made by each study participant and the amount of their effort. It is a fair measure of efficacy and also a direct and objective measure

[4] Spanish WebCAPE Computer-Adaptive Placement Exam by Jerry Larson and Kim Smith, online version Charles Bush. ©1998, 2004 Humanities Technology and Research Support Center, Brigham Young University.
[5] http://www.perpetualworks.com/webcape/overview

of efficacy. Direct, because it includes directly the effect and the effort. Objective, because the effect is measured by an independent college placement test (instead of our own test) and the effort is measured by the time recorded on computer servers (instead of self-report).

The second test used in the study was the Oral Proficiency Interview by Computer® (OPIc)[6] created by Language Testing International (LTI). LTI is the exclusive licensee of the American Council on the Teaching of Foreign languages (ACTFL). The online test is proctored and the recording of the test is reviewed and evaluated by two independent raters and an official certificate of oral proficiency is issued by ACTFL.

Table 2. OPIc Ratings

UR	Un-Ratable	**AL**	Advanced Low
NL	Novice Low	**AM**	Advanced Mid
NM	Novice Mid	**AH**	Advanced High
NH	Novice High	**S**	Superior
IL	Intermediate Low		
IM	Intermediate Mid		
IH	Intermediate High		

The specific definition of the levels are presented on the company's webpage[7].

[6] http://www.languagetesting.com/oral-proficiency-interview-by-computer-opic
[7] http://d2k4mc04236t2s.cloudfront.net/wp-content/uploads/2013/07/ACTFL-Proficiency-Guidelines-2012.pdf

3. Sample Description

The entire sample selection process is graphically represented in the Appendix, Figure A1. E-mail messages were sent out to busuu clients with an invitation to participate in the research study. If they accepted the invitation they were asked to complete the online Entry Survey with some demographic questions and questions about their knowledge of Spanish. In all 2,716 people viewed the invitation page and of those 743 successfully completed the Entry Survey. This was the initial pool of respondents in the study.

Initial Pool (N=743)

The initial pool of potential participants consisted of people from London (N=394) and New York (N=349) and about half of them (48%) were female. Interestingly, 40% of the respondents in London were female, versus almost 60% in New York. The mean age was 36 years and they were well educated: 28% had a graduate degree and 59% had a BA or some college education. Only 13% had just High School or less. The racial composition was 18% Black/African American, 13% Asian, 59% White and 10% Other Race, including multiracial categories. Most of the people were employed either full time or part-time (77%), 10% were unemployed and 13% were students.

For 78% of the initial pool English was their native language and the remainder (22%) included about 50 languages: Akan, Arabic, Azerbaijani, Cantonese, Cebuano, Chinese, Czech, Dutch, Dutch/Flemish, Filipino, French, Georgian, German, Greek, Gujarati, Haitian Creole, Hebrew, Hindi, Italian, Japanese, Korean, Lithuanian, Malay, Malayalam, Mandarin, Mauritian Creole, Nepali, Persian, Polish, Portuguese, Punjabi, Romanian, Russian, Sesotho, Shona, Sinhala, Slovak, Slovenian, Swedish, Tagalog, Tamil, Telugu, Turkish, Urdu, Uzbek, Yiddish.
Almost 97% described themselves as Novice users or Never Studied Spanish. A small proportion of them (5.2%) were of Hispanic origin and about 17% of the respondents' spouse, partner, or close friends spoke Spanish. A very small proportion (6.3%) of their parents, grandparents, or great-grandparents spoke Spanish.

The primary reason for studying Spanish was personal interest (61%), followed by business or work (17%), travel (15%), school (2%), and other reasons (5%). For other reasons the respondents mentioned: “all of the above”, “mixture”, “girlfriend/boyfriend/friend speaks Spanish”, “want more education”, “planning to move to Spain”, “to talk with Spanish family members”, “to read Don Quixote”, etc.

About 80% of the initial pool had studied a foreign language before (mostly at school or college) and about half (45%) knew a different foreign language.
About half (55%) of the participants planned to use a desktop or laptop/tablet for studying Spanish with the rest (45%) using their smartphones.

Pool of Eligible Participants (N=634)

From the Initial Pool (N=743) we excluded the following ineligible participants:

- People who were younger than 18 years of age.
- People of Hispanic origin.
- People with advanced or fluent Spanish.
- People who did not live in London or New York area.

Altogether 109 people were ineligible for this study and the final pool of eligible participants for sample selection was N=634.

The pool of eligible potential participants had a mean age of 36 years, from 18 years old to 79 years old, with 48% female users. Racial composition: 16% Black/African American, 13% Asian, 60% White, and 10% Other Race. The pool of eligible users was very well educated with only about 12% with just a High School diploma or less. About 78% were employed full time or part time, 13% were students, and 10% were unemployed. For 78% of them English was their native language and 46% of the pool knew at least one foreign language.

Initial Random Sample (N=196)

The research design determined a sample size of N=200 based on our previous studies' effect size results, drop-out rate and financial considerations. The people in the initial sample were randomly selected from the pool of eligible participants. They completed the baseline WebCAPE placement test in Spanish and the oral proficiency OPIc test (96 people in London and 100 people in New York, 4 people did not show up for the tests). Both tests were proctored.

The initial random sample had a mean age of 36 years, from 19 years old to 69 years old, with 52% female users (40% female users in London and 63% in New York). The racial composition was 20% Black/African American, 15% Asian and 55% White, and 10% Other Race. The users were very well educated with 32% with a graduate degree and 60% with some college or college degree. About 78% of them were employed either full time or part time, 13% were students, and 10% were unemployed. For 89% of them English was their native language (99% in London and 80% in New York) and almost 41% of the sample knew at least one foreign language.

People planning to use desktop/laptop/tablet made up over half the sample (58%) with the other group planning to use their smartphones. Personal interest was the primary reason (65%) for studying Spanish, followed by travel (15%), business/work (13%), school (3%) and other reason (5%). About 14% of participants' spouse/close friend and only 1% of parents/grandparents knew Spanish.

Table 3. Initial Random Sample: Age and Gender Distribution (N=196)

Age	Female (N)	Male (N)	Total (N)	Percent
18-20 years old	3	7	10	5.1
21-30 years old	45	32	77	39.3
31-40 years old	20	28	48	24.5
Over 40 years old	33	28	61	31.1
Total	**101**	**95**	**196**	**100.0**

After the selection the study participants were asked to come to New York and London offices where the initial tests were proctored. All 196 participants took the initial written online WebCAPE test and the OPIc oral proficiency test in Spanish.

The written proficiency of the initial study sample was as follows:

Table 4. Initial WebCAPE Semester Placement (N=196)

College Semester	People (N)	Percent
First	170	86.7
Second	20	10.2
Third	6	3.1
Fourth+	0	0
Total	**196**	**100.0**

The majority (87%) of the participants were evaluated as novice/beginner users of Spanish and they were placed in First Semester of Spanish. About 13% of the participants were placed in Second or Third Semester of Spanish. The mean WebCAPE score was 120 (std[8]=118) corresponding to First college semester of Spanish.

The oral proficiency of the initial sample was as follows:

Table 5. Initial Oral Proficiency (OPIc) (N=196)

Proficiency level	People (N)	Percent
1 Un-Ratable	16	8.2
2 Novice Low	133	67.9
3 Novice Mid	35	17.9
4 Novice High	6	3.1
5 Intermediate Low	4	2.0
6 Intermediate Mid	2	1.0
Total	**196**	**100.0**

The majority of people (76%) had the two lowest levels of oral proficiency and about 18% were at Novice Mid level. A handful of people (N=12) had up to Intermediate Mid level oral proficiency.

[8] Standard Deviation.

Final Study Sample (N=144)

The study continued for approximately 8 weeks, starting in February 2016 and ending in April 2016. During the study the Research Team sent weekly e-mail reminders to the participants with information detailing the amount of time they had used busuu each week.

At the end of the study we reviewed the time use of the participants. For the written proficiency test (WebCAPE) based on our previous studies (Vesselinov & Grego, 2012, 2016) the threshold was established at two hours of study. For the oral proficiency test based on the requirements by ACTFL for test-retest (2-3 months) the threshold was established at 16 hours of study. In other words, participants with at least two hours of study could take the written test and complete the study. If they had less than 16 hours of study they were not eligible to take the oral test. People with 16 or more hours of study could take both written and oral tests.

Based on these criteria from the initial sample the following people were excluded:

- People who did not satisfy the study time requirements.
- People who did not take the final tests.
- People who did not do oral practice (for the oral test only).
- People who used additional learning tools during the study.

All participants were instructed at the beginning of the study that they were allowed to use only busuu to study Spanish for the duration of the study. In the exit survey several people stated that they had regularly used other tools like other language apps, language classes, etc. and these people were excluded from the study. Other people had occasionally used internet dictionaries and similar websites and they were allowed to stay in the study.

A small portion (N=12) of the initial sample declared at the end of the study that they could not do oral practice mostly for technical reasons (e.g. microphone was not working) and these people were excluded from the final oral tests.

About 8% of the participants had less than two hours of study and they were excluded from the study. About 43% had from two hours to 15.9 hours and they were eligible to take the written proficiency test but not the oral test. Finally, about 49% had 16 hours of study or more and they were eligible to take both written and oral proficiency tests. The mean study time for the final study sample was about 18.6 hours.

The final study sample for written proficiency consisted of 144 people with at least two hours or more of busuu use and valid initial and final WebCAPE tests. The final subsample for oral proficiency was part of the 144 people sample and consisted of 61 people with about 16 hours of study or more and valid initial and final OPIc tests. The mean study time for the oral test sample was about 24 hours.

The final study sample (N=144) had a mean age of 36.6 years, from 19 years old to 69 years old, with 49.3% female users. Racial composition: 18% Black/African American, 13% Asian, 58% White and 10% of other race. The users were very well educated with 33% holding a graduate degree and 56% with BA or some college. About 76% of them were employed full time or part time, 14% were students, and 10% unemployed.

For 92% of them English was their native language and the rest included: Akan, Cebuano, Chinese, French, Italian, Malayalam, Polish, and Russian. About 40% of the sample knew at least one other foreign language (not Spanish) and they were novice users of Spanish or had never studied Spanish.

About 13% of the respondents' spouse, partner, or close friends spoke Spanish. None of their parents, grandparents, or great-grandparents spoke Spanish.
The primary reason for studying Spanish was personal interest (67%), followed by travel (17%), business or work (9%), school (3%), and other reasons (5%).

Table 6. Final Study Sample: Age and Gender Distribution (N=144)

Age	Female (N)	Male (N)	Total (N)	Total (%)
18 to 20 years old	3	6	9	6.3
21-30 years old	32	21	53	36.8
31-40 years old	12	23	35	24.3
Over 40 years old	24	23	47	32.6
Total	**71**	**73**	**144**	**100.0**

People from the final sample planned to use different devices to study Spanish with busuu. The majority of them (60%) planned to use a desktop/laptop/tablet computer with the rest (40%) planning to use smartphones. Some people used more than one device and/or different operating systems to access busuu; in the exit survey about 47% said that they had used a desktop/laptop computer, 27% had used a tablet, 37% had used an Android smartphone, and 33% had used an Apple iOS smartphone.

Final Study Sample vs Not Completed

From the initial random sample (N=196) 52 people did not complete the study for different reasons: people who did not satisfy the study time requirements; people who did not take the final tests; people who did not do oral practice (for the oral test only) and people who used additional learning tools during the study.

We compared the two groups, the final sample of 144 people and the 52 people who did not complete the study by gender, race, age, education, employment status, initial knowledge of Spanish (initial WebCAPE score and OPIc) and reason for studying Spanish. There were no statistically significant differences (at 1% error) which means that people who did not complete the study were not very different from the ones that did and they did not introduce a bias.

We also compared the sample composition for the two study sites: London and New York. There were no statistically significant differences (at 1% error) between the two sites on race,

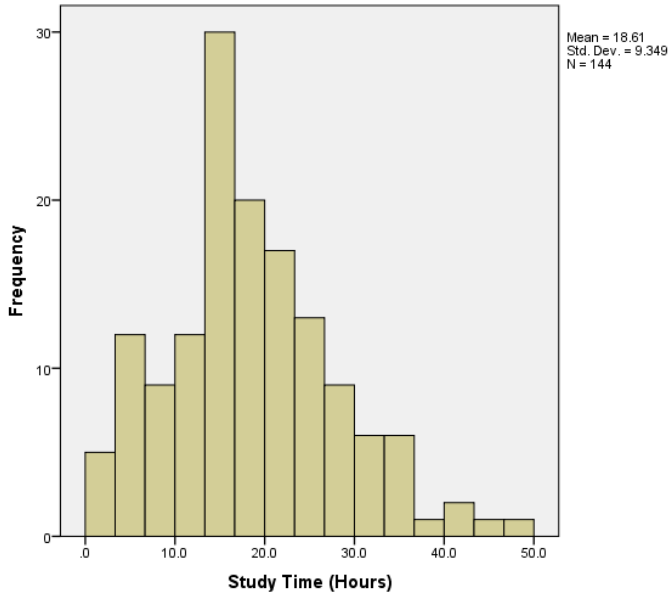
age, education, employment status, initial knowledge of Spanish (initial WebCAPE score and OPIc) and reason for studying Spanish. The London sample had significantly fewer female participants (40%) compared to 60% in New York but from our previous studies (Vesselinov & Grego, 2012, 2016) we know that gender is not a statistically significant factor in studying languages with language apps. That is why we have pooled the data from the two sites and reported the results from the combined data.

4. Language Improvement and Study Time

Study Time

The study time was measured objectively by the actual server time on a weekly basis and the time was reported to the participants regularly via e-mail in order to encourage them to keep studying. The average study time for the final study sample (N=144) was about 18.6 hours, or a little over two hours a week.

Figure 1. Study Time Distribution (N=144)



WebCAPE Test Results

All participants took a proctored initial WebCAPE test before the start of the study and then again at the end of the study. The progress or improvement was measured as the difference between the final test score and the initial one.

Table 7. Language Improvement Written Proficiency (N=144)

WebCAPE Test Points

Statistics	Initial WebCAPE	Final WebCAPE	Improvement (Final-Initial)
Mean (std)	124.8 (122.4)	269.6 (151.4)	144.8 (160.3)
Median	93.0	265.5	125.5
95% Confidence Interval[9]	104.6 – 145.0	244.6 – 294.5	118.4 – 171.2

The overall average improvement of 144.8 WebCAPE test points was statistically significant with a 95% Confidence Interval from 118 to 171 points. This also means that the improvement in the written proficiency for the final sample was statistically significant (at 5% error). Overall 84% of all participants improved their written proficiency (increased their WebCAPE score) with 95% Confidence Interval[10] of 77% to 89%.

There were 23 cases (16%) where study participants did not improve their WebCAPE result or had a lower result at the end of the study compared to their initial level. There are two plausible explanations for this fact. First, some of them were more advanced learners of Spanish (second semester) and gaining points at this higher level is generally more difficult and requires more time. Second, some of them studied irregularly with more effort and study time in the beginning of the study and less towards the end of the study. These users were **not** excluded from the sample so the results can be generalized for all types of users and not only for diligent, hardworking users who study regularly.

[9] We also bootstrapped (N=10,000) the confidence intervals but the results remained practically the same.
[10] 95% CI with Agresti-Coull correction (Agresti & Coull, 1998).

College Semester Placement

Progress can be measured by movement from one semester level to a higher semester level and the results are presented below.

Table 8. WebCAPE Semester Placement (N=144)

College Semester	Initial Test		Final Test	
	People (N)	%	People (N)	%
First	122	84.7	76	52.8
Second	17	11.8	34	23.6
Third	5	3.5	19	13.2
Fourth+			15	10.4
Total	144	100	144	100

People at First Semester level decreased from 85% to 53% and the proportion of people in Second to Fourth+ Semester level increased notably.

Oral Proficiency

The oral proficiency results are presented below.

Table 9. Oral Proficiency Placement (N=61)

Level	Initial Test		Final Test	
	People (N)	%	People (N)	%
1 Un-Ratable	6	9.8	2	3.3
2 Novice Low	30	49.2	9	14.8
3 Novice Mid	19	31.1	17	27.9
4 Novice High	2	3.3	21	34.4
5 Intermediate Low	3	4.9	7	11.5
6 Intermediate Mid	1	1.6	3	4.9
7 Intermediate High			1	1.6
8 Advanced Low			1	1.6
Total	**61**	**100**	**61**	**100**

In the oral proficiency area the results are even stronger. At the beginning of the study the truly novice users (Un-Ratable and Novice Low) were almost 60% of the sample while at the end their numbers decreased to about 18%. This is a completely different level of oral proficiency.

5. Main Results

Written Proficiency

Table 10. Written Proficiency Improvement (N=144)

Level (Semester Change)	Improved		Study Time
	People (N)	%	Mean (Hours)
-1 Negative change	2	1.4	17.4
0 Same/No Change	82	56.9	16.8
1 One Semester Up	37	25.7	21.4
2 Two Semesters Up	14	9.7	21.2
3 Three Semesters Up	9	6.3	19.6
Total	**144**	**100**	**18.6**

Overall about 42% of the participants moved up at least one semester. About 26% moved up one semester, 10% moved up two semesters and 6% moved up three semesters. About 57% stayed in the same semester they started in and two people moved down a semester. As the results indicate, the people who had invested the lowest amount of effort and study time were unsurprisingly the ones who did not improve their written proficiency in semester level.

The problem with this measure is that first, it does not account for the effort (study time) and second, moving up a semester is dependent on the exact initial level. For example, if a person has initially 269 test points (First semester), only 1 point progress is needed to move to Second semester. Another person can start with 10 points (First semester), then gain 200 points and the new level (210 points) is still First Semester.

The main efficacy measures are presented below.

Table 11. Main Result. Efficacy of busuu (N=144)

Statistics	**Efficacy Improvement per one hour of study** WebCAPE Test Points	**Time to cover the requirements for one semester of college Spanish** Hours
Mean	12.0	22.5[11]
95% Confidence Interval	8.0 – 16.0[12]	17.0 – 34.0[13]

On average busuu users will gain 12 WebCAPE test points per one hour of study with 95% Confidence Interval of 8 to 16 test points per hour.

The main measure of busuu efficacy is the improvement per one hour of study. In addition if we divide the required cut-off point (270) for WebCAPE Second Semester placement by the efficacy mean we can construct a new measure representing the time needed to cover the requirements for one college semester of Spanish. This is the one measure of efficacy that is easy to understand and given the nature of the WebCAPE placement test, can be used for comparison with other language apps.

In other words, on average, busuu users will need 22.5 hours of study to cover the requirements for one college semester of Spanish with transformed lower and upper limits of 17 hours to 34 hours of study.

[11] The threshold of 270 points divided by the mean efficacy (12 points).
[12] We also bootstrapped (N=10,000) the confidence interval but the result remained practically the same.
[13] The threshold of 270 points divided by the lower limit (8) and the upper limit (16) of the 95% CI.

Efficacy and the Initial Level of Knowledge of Spanish

Table 12. Efficacy by Initial Level of Language Ability (N=144)

Initial Level College Semester	People N	Efficacy Mean (std)
First	122	13.6 (25.7)
Second	17	2.9 (6.1)
Third	5	2.6 (3.3)
Total	**144**	**12.0 (24.0)**

The overall efficacy is 12 WebCAPE points per one hour of study but novice users of Spanish managed a bigger gain of 13.6 points per hour of study. For the second and third semester levels the improvement was more modest at about 3 points per hour.

Factors for Written Proficiency

We investigated the impact of some factors on the efficacy measure, namely age, gender, education, race, employment, device used, reason for studying Spanish, presence of people around the participant who spoke Spanish (spouse, friend, parents, grandparents, etc.), native language, and knowing another foreign language.

None of the available factors had a statistically significant effect on the efficacy. In some instances the number of cases by subgroups was too low to expect enough statistical power for the test of hypotheses.

This result can be interpreted as a positive finding because it means that the busuu app works similarly for people with different gender, age, race, employment status, native language, etc.

Oral Proficiency

Table 13. Oral Proficiency Improvement (Initial to Final Level) (N=61)

Progress in Levels	Progress		Study Time
	People (N)	%	Mean (Hours)
0 Same Level	15	24.6	20.7
1 One Level Up	27	44.3	25.3
2 Two Levels Up	15	24.6	21.2
3 Three Levels Up	3	4.9	22.7
4 Four Levels Up	0	0	n/a
5 Five Levels Up	1	1.6	47.5
Total	**61**	**100**	**23.4**

Overall 75.4% of the participants improved their oral proficiency by at least one level. The 95% Confidence Interval[14] is 63% to 85%. Almost a half (44.3%) improved by one level, a quarter (24.6%) improved by two levels and four people (6.5%) improved by more than two levels. Not surprisingly people who did not improve their oral proficiency were the people with the lowest amount of study time.

Given the study time requirements for the oral test subsample it looks as though 16 hours within two months of study, or at least two hours a week, was sufficient time for most people to achieve significant progress in their oral proficiency. The majority of people improved by one or two levels which seems like a reasonable expectation. The ACTFL requirement to have 2-3 months between test-retest looks appropriate. This study just added the additional requirement that in practice this means two months of study with at least two hours of study a week.

[14] 95% CI with Agresti-Coull correction (Agresti & Coull, 1998).

Table 14. Oral Proficiency Improvement by Initial Level (N=61)

Initial Level	People	Improved (One or more levels)
	N	% Improved
0 Un-Ratable	6	83.3
1 Novice Low	30	66.7
2 Novice Mid	19	84.2
3 Novice High	2	100
4 Intermediate Low	3	66.7
5 Intermediate Mid	1	100
Total	**61**	**75.4**

Conclusions from the oral proficiency tests were less clear than the written proficiency tests. People who started at rock bottom level as Un-Ratable and people at Novice Mid level improved the most. People at Novice Low level improved the least probably because this level can have a very diverse population; from people who literally know just a few words to people who know much more but not in a consistent way and they cannot speak Spanish in complete sentences. The more advanced levels had very few cases to make meaningful conclusions. So it looks like the oral proficiency improvement does not depend on the initial level in a clear fashion.

Factors for Oral Proficiency

We investigated the impact of some quantifiable factors on the oral proficiency measure, namely age, gender, education, race, employment, device used, reason for studying Spanish, presence of people around the participant who spoke Spanish (spouse, friend, parents, grandparents, etc.), native language, and knowing another foreign language. None of the available factors had a statistically significant effect on the oral proficiency.

As with the factors for written proficiency, this result can be interpreted as a positive finding because it means that the busuu app works the same way for people with different gender, age, race, employment status, native language, device used etc. regarding their oral proficiency.

Oral and Written Proficiency Relationship

The correspondence between the initial written proficiency and the improvement in the oral proficiency is presented below.

Table 15. Oral Proficiency Improvement by Initial WebCAPE Level (N=61)

(Predictive Relation)

Initial Written Proficiency (WebCAPE Level)	People	Improved Oral Proficiency (One or more levels)
	N	%
1 First Semester	44	72.7
2 Second Semester	13	76.9
3 Third Semester	4	100.0
Total	**61**	**75.4**

The above relationship seems to confirm the common sense logic: people who had better initial written proficiency tended to have better oral proficiency gains at the end of the study. This conclusion is in a predictive fashion since if we know the initial written proficiency we can predict the direction of the final oral gain: the higher the initial written proficiency the bigger the gain in oral proficiency. This is just a direction of the effect because it is not a statistically significant result due to the small sample size of some of the groups.

The relationship between the final written proficiency and the oral gain is presented below.

Table 16. Oral Proficiency Improvement by Final WebCAPE Level (N=61)

(Concurrent Relationship)

Final Written Proficiency (WebCAPE Level)	People	Improved Oral Proficiency (One or more levels)
	N	%
1 First Semester	20	65.0
2 Second Semester	20	80.0
3 Third Semester	12	75.0
4 Fourth+ Semester	9	88.9
Total	**61**	**75.4**

The interpretation is that people who at the end of the study still had the lowest level (First Semester) of written proficiency tended to have the lowest oral proficiency gain. Again, this is just the direction of the effect; because of the small sample size the result is not statistically significant.

The Best Results: People Who Gained in Both Written and Oral Proficiency

Overall 84% of the participants improved their written proficiency (gain in WebCAPE points) and 75% improved their oral proficiency by at least one level. The relationship between the two is presented below.

Table 17. Oral and Written Proficiency Improvement (N=61)

Improved Written Proficiency (Gain in WebCAPE score)	**Improved Oral Proficiency (One or more levels)**	
	No	Yes
No	**0 (0%)**	6 (9.8%)
Yes	15 (24.6%)	40 (65.6%)

The interpretation of this relationship is as follows: **all** people who took both the written and oral tests twice gained either in oral or written proficiency. In other words, the study shows that if a busuu user does at least two hours of study a week for two months s/he will improve in written or oral proficiency, or both. Almost 66% of the people improved both their written and oral proficiency. About 25% improved only their written proficiency and about 10% improved only their oral proficiency.

6. User Satisfaction

After the study the participants were asked for their opinion about busuu, specifically how easy it was to use, how helpful, enjoyable, and satisfactory.

Table 18. User Satisfaction (N=95)

Percent

Do you agree with the following statement?	Strongly Disagree/ Disagree	Neither Disagree nor Agree	Agree/ Strongly Agree
"busuu was easy to use"	5.3	8.4	86.3
"busuu was helpful in studying Spanish"	6.3	9.5	84.2
"I enjoyed learning Spanish with busuu"	5.3	16.8	77.9
"I am satisfied with busuu"	11.6	14.7	73.7

After two months of study, the majority of users (74% to 86%) agreed with the positive statements that: busuu was easy to use, helpful, they enjoyed learning with busuu and were satisfied with it.

In the exit survey a special question was included: "How likely are you to recommend busuu to a colleague or friend?" with 11 possible answers, from 0 "Very unlikely" to 10 "Very likely". The answers to this question were used to compute the so called Net Promoter Score (NPS). This is "a management tool that can be used to gauge the loyalty of a firm's customer relationships" (Wikipedia). It was developed by Reichheld (2003) and it categorizes users in three categories: "Promoters" (answers 9, 10), "Passives" (answers 7, 8), and "Detractors" (answers 0-6). NPS is equal to the difference between "Promoters" and "Detractors" and in general it can vary from -100 (all detractors) to + 100 (all promoters). As a rule positive NPS is good news for the company and the higher the score the better indicator for the company. From our exit survey the "Promoters" were 37.9% and the "Detractors" were 29.5% and "Passives" were 32.6%. The busuu NPS was +8.4 which is a positive result.

All of the respondents in the exit survey declared that they will continue to use busuu after the study ends.

7. Limitations of the Study

On the positive side, this was the first study for our Research Team in which the language tests were proctored (100% of the initial tests and most of the final tests) and both written and oral proficiency were evaluated with an objective measure of the study time. Neither of the tests was tailored to any specific learning tool, including busuu. On the one hand, some participants in the study complained that the tests sometimes contained words or expressions that were not part of their regular course with busuu. On the other hand, people insisted that they had learned a lot more than the tests asked for. The tests are valuable as an independent tool for evaluation which allows us to compare efficacy across different apps, however they do not provide a complete measure of the exact progress of users.

There are some limitations of the study, mostly related to the instruments and technological limitations. The WebCAPE written test measures the progress of beginner/novice users of Spanish well, but it is not suitable to measure the progress of very advanced users. Also, more study time is required for advanced users because it takes longer to achieve mastery of higher language levels. Participants who started at rock bottom as true beginners (WebCAPE score close to 0) gained much faster per study hour than people who started at the level of a second or third college semester of Spanish.

The Research team sent e-mail messages every week with individualized information about the study time for the previous week. This seemed to stimulate the study process. In normal settings when people work individually on their studies, this stimulation is not available. Many participants suggested adding a clock and time tracker to the software so they can be aware of how much time they spend studying. The average study time was about two hours of study a week but for some of the participants this was too much. The results of the study should be valid in a setting where the users study regularly for about two hours a week for two months. The study results could be generalized for studying Spanish with busuu. For other languages more studies are necessary to confirm these findings, although there is no obvious reason in the literature that the results should be markedly different.

There are few other studies with a direct objective measure of efficacy available to compare with the results of this study. More help is needed from users, investors, and analysts to require the creators of language learning apps to provide independent efficacy measures.

8. Conclusion

The busuu efficacy study is based on a final random sample of 144 people, 18 years of age or older, residing in or near London, UK or New York, US. They were not of Hispanic origin and they were novice/beginner users of Spanish.

Overall 84% of the participants improved their written proficiency (gained WebCAPE points). The main goal of measuring the efficacy of busuu was achieved with this study. The results show that, on average, one hour of study with busuu alone leads to an improvement of 12 points on the college placement test WebCAPE. There is a lot of variability of the efficacy and the 95% confidence interval is between 8 and 16 points per hour.

In other words, a busuu user would need on average 22.5 hours to complete the requirements for one college semester of Spanish. The transformed upper and lower limits are between 17 and 34 hours of study.

The main factor for the progress is the initial level of language knowledge of the participants. The novice/beginner users (First semester) gain faster, with an average of 13.6 points per one hour of study and the more advanced users (Second and Third semester) gain on average 3 points per one hour of study.

Using busuu for two months (two hours of study a week) improved the oral proficiency of 75% of the users. The 95% Confidence Interval is 63% to 85%.

There are only a handful of known studies with direct objective measures of efficacy of language learning apps. Among them the efficacy of busuu is the best so far. The creators of other language apps should be encouraged to provide efficacy measures so users and investors can make more educated choices.

9. Cited Literature

Agresti, A., B. Coull, 1998, Approximation is better than “exact” for interval estimation of binomial proportions, *American Statistician*, 52, pp. 119–126.

Reichheld, F., 2003, "One Number You Need to Grow", Harvard Business Review, 2003 December.

Vesselinov, R. & J. Grego, 2016, Efficacy of New Language App, forthcoming.

Vesselinov, R. & J. Grego, 2012, Duolingo Effectiveness Study, http://static.duolingo.com/s3/DuolingoReport_Final.pdf

Vesselinov, R., J. Grego, B. Habing, A. Lutz, 2009a, Measuring the Attitude and Motivation of Rosetta Stone ® Users, http://vesselinov.com/RV_Language_Motivation.pdf .

Vesselinov, R., J. Grego, B. Habing, A. Lutz, 2009b, Comparative Analysis of Motivation of Different Language Learning Software, http://vesselinov.com/RV_ComparativeLanguage.pdf .

Vesselinov, R., 2008, Measuring the Effectiveness of Rosetta Stone®, http://resources.rosettastone.com/CDN/us/pdfs/Measuring_the_Effectiveness_RS-5.pdf.

10. Appendix

Figure A1. Sample Selection Tree