Efficacy of a New Language App

FINAL REPORT

RESEARCH TEAM

ROUMEN VESSELINOV[1], 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

Note. The actual name of the language app tested in this study was replaced by the generic "Language App" and "LA" because the report was not officially made public.

May 2015

[1] Corresponding author.

EXECUTIVE SUMMARY

This study of efficacy of Language App (LA) was independently conducted from February to April, 2015. A random representative sample of LA users was drawn. The participants had to be at least 18 years of age, not from Hispanic origin, not living in a Spanish-speaking country and not advanced learners of Spanish.

In the beginning of the study the participants took one college placement Spanish language test and then they studied for two months using only LA. At the end of the study period they took the same test again. The improvement in language abilities was measured as the difference between the final and the initial language test results. The efficacy of LA was measured as language improvement per one hour of study.

MAIN RESULTS

- The efficacy of LA is a gain of about 11 test points per one hour of study.
- Beginner users of Spanish gain on average about 18 test points per one hour of study.
- More advanced users gain on average 4 to 6 test points per one hour of study.
- LA users would need on average 25 hours of study to cover the requirements for one college semester of Spanish (95% confidence interval: 18 to 41 hours).
- Almost half (46%) of the study participants moved up at least one college semester level; 11% moved up two semesters and 3% moved up three semesters.
- The majority of the users (81% to 94%) thought that LA was easy to use, helpful, and enjoyable and they were satisfied with it.

There are only a handful of known studies with direct objective measure of efficacy of language learning software packages. Among them the efficacy of LA is the best so far. The creators of these products should be encouraged to provide efficacy measures so consumers can make more educated choices.

Contents

Introduction

Nowadays learning new languages with the help of language learning software or applications is becoming more and more popular. There is a growing interest in evaluating the efficacy (or effectiveness) of the language learning software packages or applications. 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 four studies attempting to directly evaluate the efficacy, attitude and motivation of some popular language learning software packages, namely Rosetta Stone®, Aurolog® and Berlitz® (Vesselinov 2008, Vesselinov et al. 2009a, 2009b), and Duolingo (Vesselinov & Grego, 2012).

With this study, we are trying to evaluate the efficacy of a newly developed language software product, LA.

This study was funded by LA but the data collection and the analysis were done independently by the Research team.

Research Design

LA provided to the Research team the e-mail addresses of their new subscribers and we drew a representative sample based on the pool of eligible users. The following list of criteria was used in the selection of users who were:

- Willing to participate in the study;
- Studying Spanish as a foreign language;
- At least 18 years of age;
- Not of Hispanic origin;
- Not living in a Spanish-speaking country;
- Not advanced users of Spanish.

The last requirement was due to the fact that the language placement test used in the study has placement in college Semester 4+ as its highest evaluation group and it has limited abilities for the very advanced users.

The LA users in the initial pool were from all over the world (five continents) including some people living in a Spanish-speaking country. Learning Spanish in a Spanish-speaking

country has many advantages and it is not a fair evaluation of LA. Therefore, these users were excluded from the sample.

The recommended goal for the participants in the study was to use LA for at least 16 hours during the two-month study, or two hours per week. We knew in advance that this recommendation would not be feasible for some participants. For this study we imposed a threshold of at least two hours of study. 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.

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 software packages. The length of the study was approximately 8 weeks and it was conducted between the months of February and April of 2015. People who successfully completed the study were given a lifetime free subscription to the Premium Edition of LA. At the time of the study this edition was not yet developed. 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[2] (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 maintained by the Perpetual Technology Group. A more detailed description of the test can be found at their website[3].

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+

[2] Spanish WebCAPE Computer-Adaptive Placement Exam by Jerry Larson and Kim Smith, WWWeb version Charles Bush. ©1998, 2004 Humanities Technology and Research Support Center, Brigham Young University

[3] http://www.perpetualworks.com/webcape/overview

The measure of Efficacy for this study was defined as follows:

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

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 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 the computer servers of the software (instead of self-report).

Sample Description

The entire sample selection process is graphically represented in the Appendix, Figure A1.

The Research team received a list with the LA users' e-mails and sent an invitation to participate in the study to all of them. If they accepted the invitation they were asked to complete the Entry survey with some demographics[4] and questions about their knowledge of Spanish. In all 584 people viewed the invitation page and of them 380 successfully completed the Entry survey. This was the initial pool of respondents in the study.

- **Initial Pool (N=380)**

The initial pool of potential participants consisted of people from five continents: North America, South America, Europe, Asia and Australia, and from 18 countries: United Arab Emirates, Australia, Canada, France, UK, Hong Kong, Iceland, Israel, India, Japan, Nepal, Taiwan, USA, and some Spanish-speaking countries: Costa Rica, Spain, Honduras, Panama, Peru. Most of the users in the Spanish-speaking countries were US expatriates working or living there. The users from the US were 285 or 75% of the initial pool and they came from about 40 US states (see Appendix, Table A2).

The initial pool of people interested in studying Spanish had a mean age of 39.3 years, from 6 years old to 78 years old, with 64.2% female users. The initial pool of users was very well educated with 71.3% holding a college degree or graduate degree. About 67% of them were

[4] Initially the Entry survey included a race category variable but after objections from users outside the U.S. this variable was removed from the survey.

employed full time or part time, 13.4% were students, and 7.4% were unemployed and the rest declared other type of employment (2.1%) or refused to answer (10.5%).

For 91.6% of them, English was their native language and the rest (8.4%) included: Arabic, Cantonese, Farsi, French, German, Greek, Albanian, Hebrew, Hindi, Korean, Macedonian, Malay, Polish, Romanian and Thai. Almost 30% of the pool knew at least one foreign language (not Spanish).

Almost 98% described themselves as Novice/Beginner to Intermediate user of Spanish. A small proportion of them (7.6%) were of Hispanic origin and about a quarter of the respondents' spouse, partner, or close friends spoke Spanish. A small proportion (9.7%) of their parents, grandparents, or great-grandparents spoke Spanish.

The primary reason for studying Spanish was personal interest (60.0%), followed by business or work (14.7%), travel (12.9%), school (2.6%), and other reasons (9.7%). For other reasons the respondents mentioned: "all of the above", "expatriate", "girlfriend/boyfriend speaks Spanish", "daughter learning Spanish in school", "live/work in Spanish-speaking country", "to talk with my Spanish family members", "the future belongs to bilingual…", etc.

- **Pool of Eligible Participants (N=326)**

From the Initial Pool (N=380) we excluded the following ineligible participants:

1. People who were younger than 18 years of age.
2. People of Hispanic origin.
3. People with advanced or fluent Spanish.
4. People who lived in a Spanish-speaking country.

Altogether 54 people were ineligible for this study and the final pool of eligible participants for sample selection was N=326.

The pool of eligible potential participants had a mean age of 40.2 years, from 18 years old to 78 years old, with 63.5% female users. The eligible pool of users was very well educated with about 73% holding a college degree or graduate degree. About 67% of them were employed full time or part time, 12% were students, and 8% were unemployed and the rest declared other type of employment (2%) or refused to answer (11%). For 91.1% of them English was their native language and almost 32% of the pool knew at least one foreign language.
The respondents were geographically from 13 countries (see Appendix, Tables A1 and A2).

-

Initial Random Sample (N=231)

The people in the initial sample were randomly selected from the pool of eligible participants. Originally 231 people were selected and they completed the baseline WebCAPE placement test in Spanish.

The initial random sample participants had a mean age of 40.5 years, from 19 years old to 78 years old, with 65.8% female users. The users were very well educated with about 75% holding a college degree or graduate degree. About 65% of them were employed full time or part time, 11% were students, and 9% were unemployed and the rest declared other type of employment (3%) or refused to answer (13%). For 91.3% of them English was their native language and almost 33% of the sample knew at least one foreign language.

The respondents were geographically from 9 countries (see Appendix, Tables A1 & A2).

Table 2. Initial Random Sample: Age and Gender Distribution (N=231)

Age	Female (N)	Male (N)	Total (N)	Percent
18-20 years old	5	0	5	2.2
21-30 years old	36	14	50	21.6
31-40 years old	46	21	67	29.0
Over 40 years old	65	44	109	47.2
Total	**152**	**79**	**231**	**100.0**

After the selection the study participants were asked to go online and complete the first WebCAPE placement test in Spanish.

Table 3. Initial WebCAPE Semester Placement (N=231).

College Semester	People (N)	Percent
First	125	54.1
Second	47	20.3
Third	37	16.0
Fourth+*	22	9.5
Total	**231**	**100.0**

* People who scored Fourth+ semester were excluded from the study.

The mean WebCAPE score was 240.2 (Median=254) corresponding to First college semester of Spanish.

Table 4. Initial WebCAPE Placement Test Statistics (N=231).

Statistics	WebCAPE Test Points
Mean (std)	240.2 (146.8)
Median	254.0
Min	0
Max	582

• Final Study Sample (N=101)

The study continued for 8 weeks, starting in February 2015 and ending in April 2015. During the study, the Research team sent weekly e-mail reminders to the participants with information detailing the amount of time they had used LA each week.

At the end of the study we reviewed the time use of the participants. The initial target for this study was at least 16 hours of use for the two months of study. About 10% of the initial sample did have 16 hours or more of study. The lowest threshold for inclusion in the study was defined as about 2 hours. People who had studied Spanish for less than 2 hours for the whole period of two months were considered not seriously studying and they did not complete the study. At the end 114 people completed the study and took the final WebCAPE test. Of them 13 were eventually excluded from the study because in addition to LA they have used other tools like college courses or other language learning software.

The question about using additional help or additional tools during the study was asked in the exit survey as a way to confirm that LA was the only tool used for studying Spanish. A few participants said that they occasionally used some web tools for additional information like dictionaries, translation sites, watched some Spanish videos etc. and they remained in the sample.

The final study sample consisted of 101 people with about 2 hours or more of LA use and valid initial and final WebCAPE tests. They were people 18 years of age and older, not from Hispanic origin, initially not advanced users of Spanish and not living in a Spanish-speaking country.

The final study sample participants were from four continents: North America, Europe, Asia and Australia, and from 6 countries: United Arab Emirates, Australia, Canada, UK, Israel, and the US. There were 81 users from the US, or 80% of the final study sample and they came from 30 US states (see Appendix, Table A2).

The final study sample had a mean age of 42.6 years, from 20 years old to 78 years old, with 66.3% female users. The users were very well educated with 75.3% holding a college degree or graduate degree, 7.9% with a high school diploma or less, and 16.8% with some college (but did not graduate). About 67% of them were employed full time or part time, 6% were students, and 8% were unemployed and the rest declared other type of employment (3%) or refused to answer (16%).

For 92% of them English was their native language and the rest included: Arabic, Farsi, French, German, Hebrew, Polish, and Romanian. Almost 28% of the sample knew at least one foreign language (not Spanish).

About 73% of the participants in the beginning of the study described themselves as a Novice/Beginner and 27% as an Intermediate user of Spanish. About 20% of the respondents' spouse, partner, or close friends spoke Spanish. A small proportion (2%) of their parents, grandparents, or great-grandparents spoke Spanish.

The primary reason for studying Spanish was personal interest (55%), followed by travel (20%), business or work (13%), school (2%), and other reasons (10%). For other reasons the respondents mentioned: "to speak with my in-laws", "girlfriend/boyfriend speaks Spanish", "live/work in a Spanish-speaking country in the near future", "to talk with my Spanish family members", "live in a state with a large Hispanic population", etc.

Table 5. Final Study Sample: Age and Gender Distribution (N=101).

Age	Female (N)	Male (N)	Total (N)	Total (%)
18 to 20 years old	1	0	1	1.0
21-30 years old	8	6	14	13.9
31-40 years old	22	8	30	29.7
Over 40 years old	36	20	56	55.4
Total	**67**	**34**	**101**	**100.0**

People from the final work sample used different devices to study Spanish with LA. The majority of them (79.4%) used desktop or laptop computer, next came the tablets (13.4%) and smartphones (7.2%).

Study Sample vs Not Completed

From the initial random sample (N=321) 133 people did not complete the study for different reasons: some were too advanced Spanish users with very high initial WebCAPE scores; some used additional tools during the study; some had less than two hours of study and some did not complete the final WebCAPE test. In that regard the group is not comprised of typical drop-outs. Regardless, since they did not complete the study the question is whether these people were statistically different than the final work sample (N=101). We compared the two groups by gender, age, education, employment status, initial knowledge of Spanish (initial WebCAPE score) and reason for studying Spanish. The only significant difference (p=.026) was for age: people who successfully completed the study were slightly older than the people who did not complete the study (42.5 vs 39 years old). Most importantly there was no statistically significant difference in their initial knowledge of Spanish.

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.

Table 6. Structure of the Study Time (N=101).

	Study Time	Percent of Study Time Spent for:		
Statistics	**Hours**	**Song Lessons**	**Course Lessons**	**Regular Lessons**
Mean (std)	10.7 (10.9)	4 (11.4)	9.6 (17.3)	86.4 (21.1)
Median	7.8	0	1.6	95.0
Min	2	0	0	3.3
Max	78.7	88.9	96.7	100.0

Overall 48.5% of the participants have tried Song Lessons at least once and 55.4% have tried Course Lessons at least once and everybody had tried the Regular Lessons. There were people who relied almost entirely on the Regular Lessons or Course Lessons, while other people preferred Song Lessons reaching almost 90% of their study time.

The average study time was about 10.7 hours with 2 hours as the lowest and 78.7 hours as the highest.

Figure 1. Study Time Distribution (N=101).

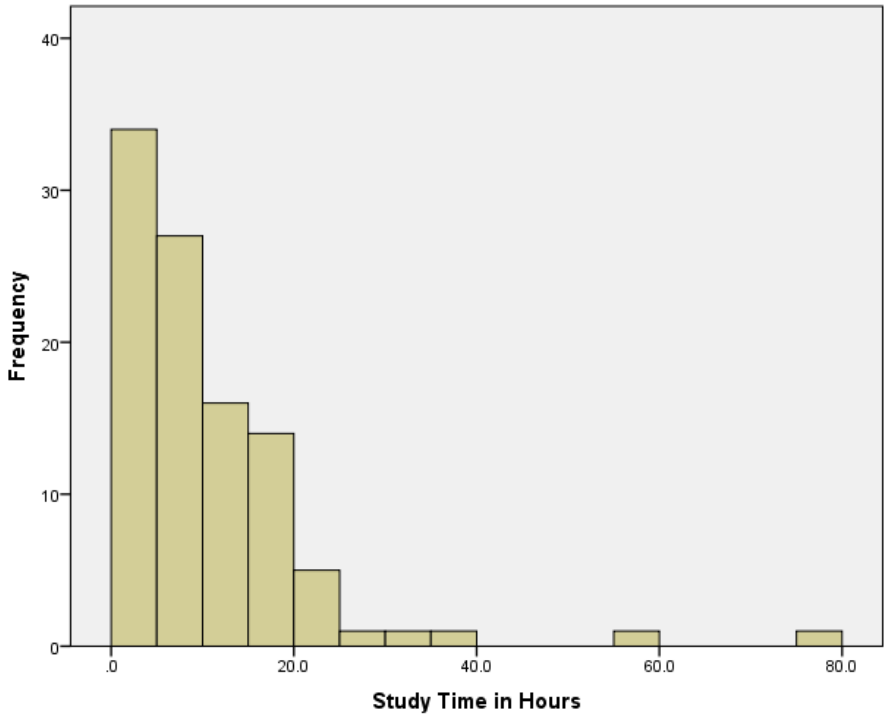


Table 7. Study Time Frequencies (N=101).

People	**Study Time (Hours)**					
	2-3	3.01-5	5.01-7	7.01-10	10.01-16	> 16
Number	12	22	11	16	22	18
Percent	11.9	21.8	10.9	15.8	21.8	17.8

- **WebCAPE Test Results**

All participants were asked to take the 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 8. Language Improvement (N=101).

WebCAPE Test Points

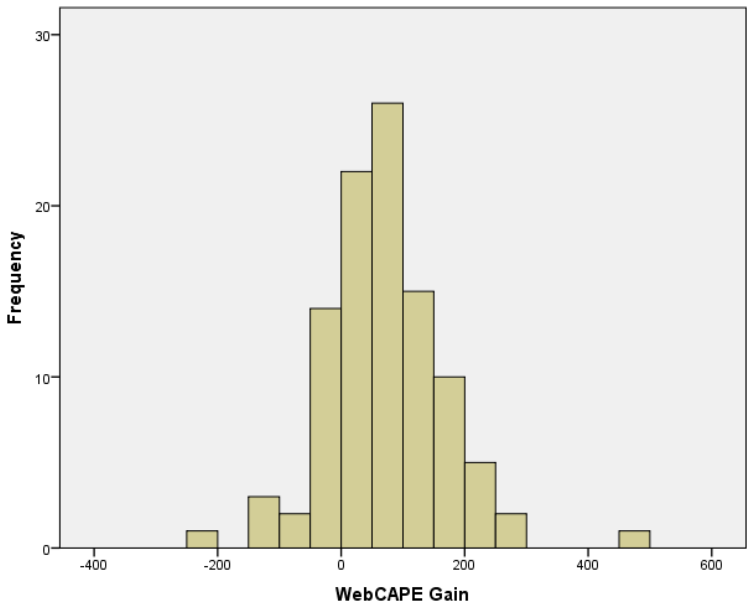
Statistics	**Initial WebCAPE**	**Final WebCAPE**	**Improvement (Final-Initial)**
Mean (std)	243.4 (117.6)	314.4 (113.8)	71.0 (96.9)
Median	278	341	61.0
Min	0	0	-239
Max	418	543	467
95% confidence interval*	220.2 – 266.6	291.9 – 336.9	51.9 – 90.1

* We also bootstrapped (N=10,000) the confidence intervals but the results remained practically the same.

The overall improvement of 71 WebCAPE test points was statistically significant with a 95% confidence interval from 51.9 to 90.1 points.

There were 21 cases where study participants did not improve their 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, most of them were more advanced learners of Spanish, initially placed in second or third semester and gaining points at this higher level is generally more difficult and requires more time. Therefore if you are an advanced user and you do not spend enough time studying, the results may not be satisfactory. Second, many of them studied irregularly with more efforts and study time in the beginning of the study and then big gaps without any. These users were not excluded from the sample so the results can be generalized for all types of users not only for diligent, hardworking and regularly studying users but also for people who study not very regularly. The biggest gain was by a study participant who started with initial WebCAPE score of 0 and after about 10 hours of study reached 467 points.

Figure 2. Language Improvement: WebCAPE Gain in Test Points (N=101).



Placement for four semester college course.

The progress here can be measured by movement from one semester level to a higher semester level. Overall 46% of the participants moved up at least one semester. About 11%

moved up two semesters and 3% moved up three semesters. About 45% stayed in the same semester they started in and 10% moved down a semester.

Table 9. WebCAPE Semester Placement (N=101).

College Semester	**Initial Test**	**Final Test**
	People	People
	Percent (N)	Percent (N)
First	47.5 (48)	34.7 (35)
Second	32.7 (33)	23.8 (24)
Third	19.8 (20)	24.8 (25)
Fourth+		16.8 (17)
Total	100 (101)	100 (101)

Table 10. WebCAPE Semester Placement Initial vs Final (N=101).

People (N)

Initial Placement (Semester)	**Final Placement (Semester)**				
	First	Second	Third	Fourth+	**Total**
First	28	11	6	3	48
Second	6	10	12	5	33
Third	1	3	7	9	20
Total	35	24	25	17	101

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 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 level (First semester), gain 200 points and the new level (210 points) is still First semester.

Main Results

• Efficacy of LA

The WebCAPE results alone cannot give a clear picture about the efficacy of the language learning software 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:

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

Or, Efficacy=Improvement per one hour of study.

Table 11. Main Result. Efficacy of LA (N=101).

Statistics	**Efficacy** WebCAPE Test Points	**Time to Cover One Semester of College Spanish** Hours
Mean	10.8	25.0
95% confidence interval*	6.6 – 15.0	18.0 – 41.0

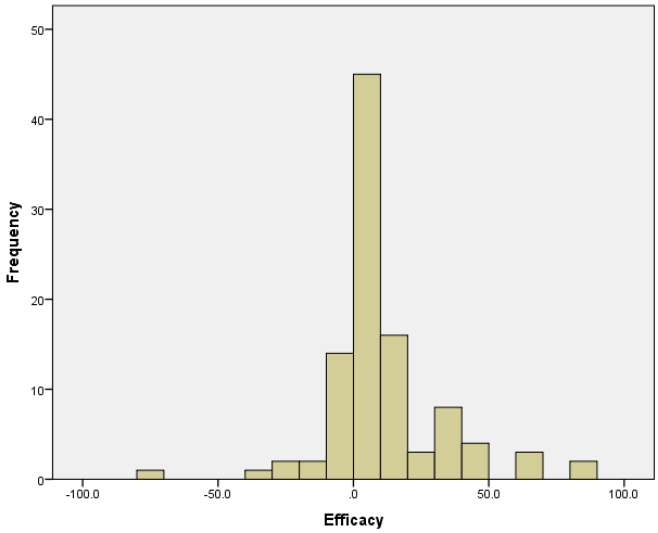
* We also bootstrapped (N=10,000) the confidence intervals but the results remained practically the same.

The maximum improvement achieved in the study was a participant with 82.4 points per hour of study with a total of about two hours of study. The worst case in the study was a participant with 74.3 points decrease per hour of study with total of about 3 hours of study.

On average LA users will gain 10.8 test points per one hour of study with 95% confidence interval of 6.6 to 15.0 test points per hour.

If we divide the required cut-off points (270) for WebCAPE second semester placement we can construct a new measure representing the time needed to cover the requirements for one semester. Thus, on average LA users will need 25 hours of study to cover the requirements for one college semester of Spanish with a 95% confidence interval from 18 hours to 41 hours of study.

Figure 3. LA Efficacy Distribution (N=101).



Efficacy and the Initial Level of Knowledge of Spanish

Table 12. Efficacy by Initial Level of Language Ability (N=101).

Initial Level College Semester	**People** N	**Efficacy** Mean (Std)
First*	48	17.6 (24.0)
Second	33	4.0 (19.8)
Third	20	5.7 (9.5)
Total	**101**	**10.8 (21.4)**

* The improvement for the first semester group was statistically different from second and third semester (t-test with Tukey HSD correction for multiple comparisons).

The overall efficacy was 10.8 WebCAPE points per one hour of study but less advanced users with an initial level of First college semester managed a bigger efficacy of 17.6 points. For the second and third semester levels the improvement was more modest from 4 to 6 points per hour.

- **Efficacy Factors**

We investigated the impact of some quantifiable factors on the efficacy measure. All but one had no 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. We are reporting the direction of some effects in case the effect is real and can become significant with larger samples in future studies.

The following factors **did not** have statistically significant effect on LA efficacy but some effects' directions are reported.

- Age. Effect direction: people over 40 years of age performed better that the younger groups;
- Gender;
- Device used to study with LA. Effect direction: people who used a smartphone as their device during the study performed worse than people working with a desktop/laptop or tablet;
- The reason for studying Spanish;
- The presence of people around the participant who spoke Spanish (spouse, friend, grandparents, etc.);
- Native language;
- Knowing other foreign languages. Effect direction: people who knew other foreign languages performed a little better than those who did not;
- The location (US vs Not US). Effect direction: participants residing in the US performed a little better.
- Education. Effect direction: people with a college degree or graduate degree had the highest efficacy;
- Employment. Effect direction: participants who were students or employed part-time had higher efficacy;
- Using different types of lessons available with LA;

The only statistically **significant** factor for efficacy was the initial level of knowledge of Spanish. Beginner/novice participants who initially placed in the First college semester had the highest efficacy and gained on average 17.6 points per one hour of study. Participants in the second or third semester gained modestly on average 4 to 6 points per one hour of study.

User Satisfaction

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

Table 13. Users Satisfaction (N=97).

Percent

Do you agree with the following statement?	Strongly Disagree/ Disagree	Neither Disagree nor Agree	Agree/ Strongly Agree
"LA was easy to use"	4.1	3.1	92.8
"LA was helpful in studying Spanish"	4.1	2.1	93.8
"I enjoyed learning Spanish with LA"	3.1	4.1	92.8
"I am satisfied with LA"	8.2	10.3	81.4

The majority of the users (81% to 94%) agreed with the positive statements that after two months of study with LA they confirmed that it was easy to use, helpful, and they enjoyed learning with LA and were satisfied with it.

In the exit survey a special question was included: "How likely are you to recommend LA 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 (N=97) the "Promoters" were 50.5% and the "Detractors" were 9.3% and "Passives" were 40.2%. The LA NPS was +41.2 which is a strong result.

Almost all (97.9%) participants in the exit survey declared that they will continue to use LA after the study ends.

Comparison with Previous Studies

Many users, investors, analysts and professionals are interested in comparisons of the existing language learning software products. Here we are discussing three options.

Study 1. Rosetta Stone® Effectiveness Study (Vesselinov, 2008)

For this study, like the LA study, the measure for effectiveness was based on the WebCAPE test and the time used by the participants. However, in 2008 objective records of the Rosetta Stone® usage time were not available because the program was a standalone version installed from CDs on home computers. User self-reported study time was used instead. In this sense the effectiveness measure was direct but subjective because the self-reported time was found to be inflated and inaccurate and the rate of inaccuracy was not known. This is the main reason the current LA results and Duolingo, 2012 results cannot be directly compared to this study's results. Also the version tested was the 2008 version of the Rosetta Stone® product and in this area seven years of development can dramatically change the performance of a product.

What is needed is a new study of Rosetta Stone® with a direct and objective measure for efficacy/effectiveness for the current version of this product.

Study 2. Duolingo Effectiveness Study (Vesselinov & Grego, 2012)

This study is directly comparable with the current LA study with minor stipulations and differences. For the Duolingo study all participants were native speakers of English and residing in the US. For the LA study a small portion (8%) of the study participants were not native speakers of English and about 20% lived outside the US. But the current study showed no statistically significant differences in the efficacy for these factors (native language and location) so comparing the two studies is possible. Lastly, the Duolingo study reflects the 2012 version of the product.

The point estimate of efficacy, transformed into hours of study, shows that LA requires on average 25 hours of study to cover the requirements for one college semester of Spanish while Duolingo requires 34 hours. But the two 95% confidence intervals are very wide, Duolingo: 26-49 hours and LA: 18-41 hours, so the difference between the two measures is not statistically significant.

Study 3. One Standard College Semester of Spanish

After the 2012 Duolingo study there were some attempts in the press to compare the results to a standard one semester of college Spanish. One semester usually implies 15 weeks with two classes (75 min each), or roughly 40 hours per semester. If we include seminars or additional requirements and assignments the total time will be about 60 hours.

Comparison between this study and the Duolingo study with one standard college semester of Spanish is not scientifically sound for two reasons.

First, progress or success in college is determined usually with one midterm exam and one final exam, plus some form of testing for speaking abilities and homework. WebCAPE or other online tests are only used as initial placement tests and not for determining course grades. So the progress is measured very differently in college and in the two existing studies.

Second, the study time is not 60 hours for everybody as the course plan says. If we use the same definition of study time in the current LA study everybody should get 16 hours as it was planned. But we know that this study time is inflated because we have objective data showing that there are many people with two hours of study, three hours, etc. and the average time is much lower than 16 hours.

If there is a need to compare to one college semester of Spanish, a new study can be designed to satisfy the efficacy definition. To the best of our knowledge such a study has not been done yet and until then the two sets of results cannot be compared scientifically.

Limitations of the Study

The progress in language abilities is based solely on an online college language placement test and does not include listening comprehension or speaking proficiency evaluation. The test is not tailored to any specific learning tool, including LA. Some participants in the study complained that the test asks for words that were not part of their regular course with LA and that they have learnt a lot more that the test does not ask for. The test is valuable as an independent tool for evaluation which allows us to compare efficacy across different tools but it does not measure all the progress of the users. It would be desirable to include as an assessment instrument other more sophisticated language learning tests, including speaking tests which unfortunately are very expensive for research purposes. One possible candidate is the Oral Proficiency Interview by Computer® (OPIc)[5] licensed by the American Council for the Teaching of Foreign Languages (ACTFL).

The design of the study and the independent evaluation test measured the progress of beginner/novice users of Spanish but are not suitable to measure the progress of very advanced users. Also, more study time may be required for advanced users because their progress is

[5] ACTFL website http://www.languagetesting.com/

slower. It seems that the efficacy has diminishing returns with placement. Participants who started at the rock bottom as true beginners (WebCAPE score close to 0) gained much faster (17.6 points per study hour) than people who started at the level of second or third college semester of Spanish (4 to 6 points per study hour).

The Research team sent e-mail messages every week with 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 have asked for adding a clock and time tracker to the software so they can be aware of how much time they spend studying. The median study time was about 8 hours which means roughly one hour of study a week. Our target time was two hours per week but this obviously was too much for many of the participants. Without the weekly regular report on study time, people (on average) would have spent less time studying.

The overall sample size of 101 participants gives sufficient statistical power to generalize the overall results of this study. But the subsamples of Second and Third semester groups are not large enough to do a separate analysis for these two groups. With larger samples for these two subgroups more detailed and separate analysis could be done.

The study results could be generalized for studying Spanish with LA. For other languages more studies are necessary to confirm these findings.

There are not many other studies with a direct objective measure of efficacy available to compare with this study's results. More help is needed from users, investors, and analysts to require the creators of language learning software to provide efficacy measures. Including the efficacy information will allow consumers to make a more educated choice.

Conclusion

This study of LA efficacy is based on a study sample consisting of 101 people, 18 years of age or older, mostly English native speakers (92%) and mostly residing in the US (80%). They were not of Hispanic origin and did not live in a Spanish-speaking country. Participants also had to spend at least two hours of study with LA.

The main goal of measuring the efficacy of LA was achieved with this study. The results show that on average, one hour of study with LA alone brings about progress of 10.8 points on the college placement test WebCAPE. There is a lot of variability of the efficacy and the 95% confidence interval is between 6.6 and 15 points.

We used the efficacy estimates and the WebCAPE college semester placement cut-off points (270 points) to create an estimate of time needed to complete the requirements. A LA user would need on average 25 hours to complete the requirements for one college semester of Spanish. The 95% confidence interval for this measure is between 18 and 41 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 17.6 points per one hour of study and the more advanced users (Second and Third semester) gain on average 4 to 6 points per one hour of study.

There are only a handful of known studies with a direct objective measure of efficacy of language learning software packages. Among them the efficacy of LA is the best so far. The creators of these products should be encouraged to provide efficacy measures so consumers can make more educated choices.

Cited Literature

Vesselinov, R. & J. Grego, 2012, Duolingo Effectiveness Study, Final Report, 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, Final Report, manuscript available through Rosetta Stone®.

Vesselinov, R., J. Grego, B. Habing, A. Lutz, 2009b, Comparative Analysis of Motivation of Different Language Learning Software, Final Report, manuscript available through Rosetta Stone®.

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

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

Appendix

Figure A1. Sample Selection Tree

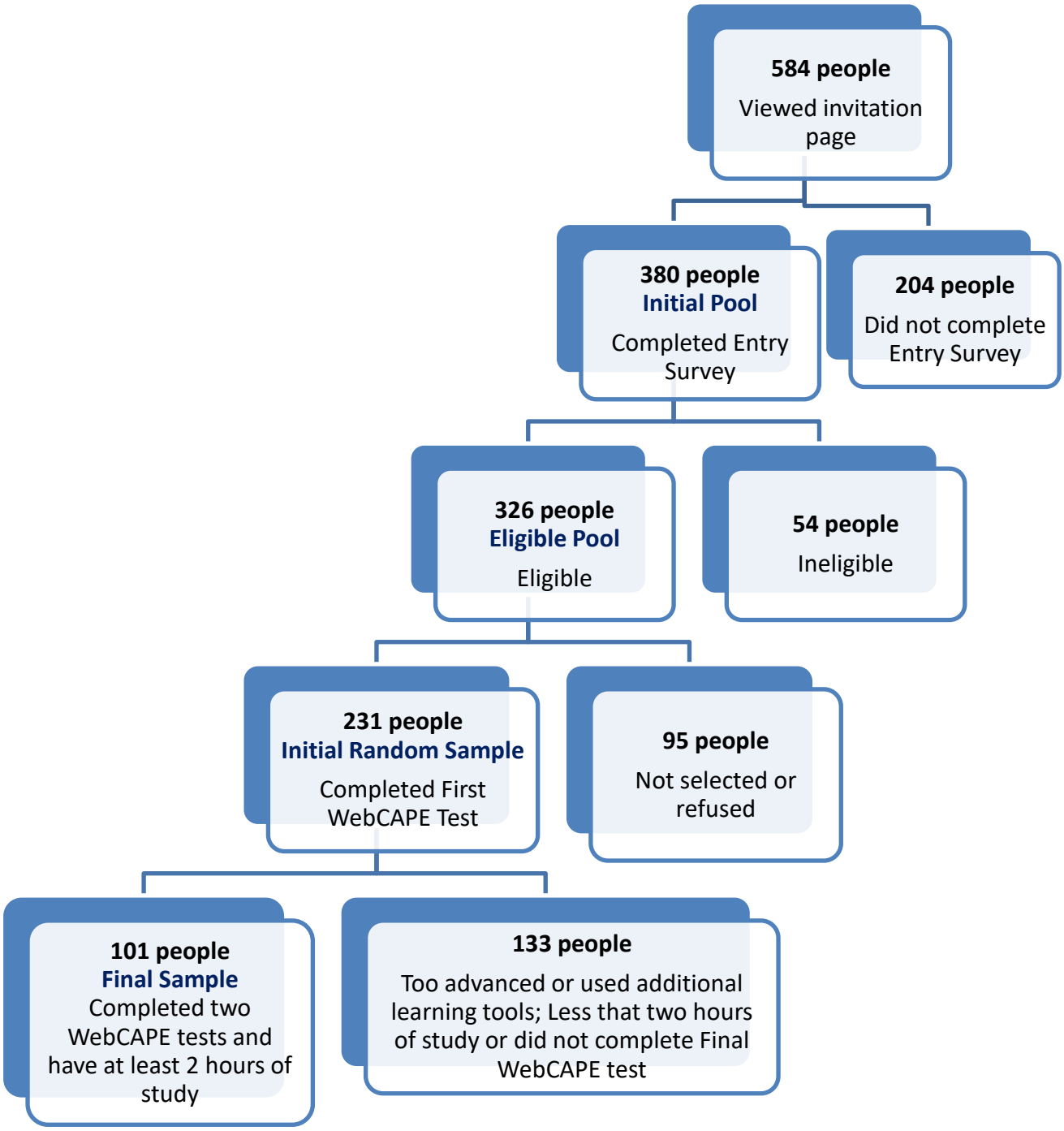


Table A1. Geographic Distribution Countries (Number of people)

	State	Country Code	Initial Pool	Eligible Pool	Initial Sample	Final Sample
1	United Arab Emirates	AE	1	1	1	1
2	Australia	AU	11	10	9	5
3	Canada	CA	16	16	11	3
4	France	FR	1	1		
5	United Kingdom	GB	37	34	20	9
6	Hong Kong	HK	3	2	2	
7	Ireland	IE	3	3	1	
8	Israel	IL	7	7	4	2
9	India	IN	1	1		
10	Japan	JP	1	1	1	
11	Nepal	NP	1	1		
12	Taiwan	TW	1	1		
13	USA	US	285	248	182	81
	Spanish-Speaking Countries					
14	Costa Rica	CR	2			
15	Spain	ES	5			
16	Honduras	HN	1			
17	Panama	PA	2			
18	Peru	PE	2			
Total			**380**	**326**	**231**	**101**

Table A2. Geographic Distribution: US States (Number of people)

	State	ST	Initial Pool	Eligible Pool	Initial Sample	Final Sample
1	Alabama	AL				
2	Alaska	AK				
3	Arizona	AZ	8	7	7	3
4	Arkansas	AR				
5	California	CA	48	42	32	10
6	Colorado	CO	3	3	2	1
7	Connecticut	CT	10	8	6	3
9	Delaware	DE	1	1		
10	Florida	FL	17	17	13	6
11	Georgia	GA	10	9	6	2
12	Idaho	ID	2	2	2	2
13	Illinois	IL	8	6	6	3
14	Indiana	IN	6	5	4	3
15	Iowa	IA	4	3	2	
16	Kansas	KS	3	3	2	1
17	Kentucky	KY	3	3	2	2
18	Louisiana	LA	3	3	1	1
19	Maine	ME				
20	Maryland	MD	7	6	4	
21	Massachusetts	MA	6	3	1	
22	Michigan	MI	5	3	3	1
23	Minnesota	MN	6	6	4	1
24	Mississippi	MS	2	2	1	1
25	Missouri	MO	1	1	1	1
26	Montana	MT				
27	Nebraska	NE	3	3	3	2
28	Nevada	NV				
29	New Hampshire	NH	1			
30	New Jersey	NJ	3	3	3	2
31	New Mexico	NM	1	1	1	1
32	New York	NY	18	17	11	5
33	North Carolina	NC	11	10	9	5
34	North Dakota	ND				
35	Ohio	OH	11	10	9	3

Table A2 Continued

	State	ST	Initial Pool	Eligible Pool	Initial Sample	Final Sample
36	Oklahoma	OK				
37	Oregon	OR	3	2	1	
38	Pennsylvania	PA	10	8	4	1
39	Rhode Island	RI	1	1	1	1
40	South Carolina	SC	1	1	1	
41	South Dakota	SD				
42	Tennessee	TN	5	5	3	2
43	Texas	TX	22	20	13	6
44	Utah	UT	3	2	2	
45	Vermont	VT	1	1		
46	Virginia	VA	12	9	8	
47	Washington	WA	11	10	6	3
48	West Virginia	WV	2	2		1
49	Wisconsin	WI	4	4	4	1
50	Wyoming	WY				
	District of Columbia	DC	1			
	Unknown state (but US)		8	6	4	7
Total		**USA**	**285**	**248**	**182**	**81**