The Babbel 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

September 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 June 2016 to August 2016. The study was based on a random representative sample of 325 Babbel users. The participants took one college placement Spanish language test, then studied Spanish with Babbel for two months and 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 Babbel was measured as language proficiency improvement per one hour of study.

MAIN RESULTS

Babbel Language Proficiency Gain:

- Overall 92% of the participants improved their language proficiency.
- Babbel users need on average 21 hours of study in a two-month period to cover the requirements for one college semester of Spanish.
- Truly novice users with no knowledge of Spanish need on average 15 hours of study in a two-month period to cover the requirements for one college semester of Spanish.

SUPPLEMENTARY RESULTS

Language Proficiency:

- The efficacy of Babbel is a gain of about 12.7 test points per one hour of study. For truly novice users with no knowledge of Spanish the gain is 18 points and for more advanced users it is about 10 points per hour of study.
- About 45% of the participants moved up at least one college semester level. Overall, 27% moved up one semester, 13% - two semesters, and 5% - three semesters.

User Satisfaction:

- The majority of users thought that Babbel was easy to use (95%), helpful (95%), enjoyable (92%), and they were satisfied with it (94%).
- Babbel received a large positive Net Promoter Score of +59.3 from the users.
- Babbel efficacy was not affected by gender, age, education, native language, device used, etc.

CONTENTS

Introduction

There is a growing interest in evaluating the efficacy (or effectiveness) of computer assisted language instruction applications, or language learning apps for short. New users, investors, analysts and academics are eager to learn what they can expect to gain by using a particular application and which one is the 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®, Berlitz®, Duolingo®, and Busuu® (Vesselinov 2008, Vesselinov et al. 2009a, 2009b, Vesselinov & Grego, 2012, 2015 and 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, researchers. etc. All of them need help to decide which language app they should use. Other things being equal (e.g. price, appearance, ease to use, etc.) they need 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: Babbel[3]. The company was founded in 2007. According to Babbel, it is the world's first language learning app and offers to be the shortest path to a real-life conversation. It is available on iOS, Android and Web. Babbel has more than one million paying subscribers for its premium subscription based service. With a world-class method at its core, Babbel offers courses to learn up to 14 different languages and teaches conversational skills learners can use straight away and with confidence.

Babbel main features are as follows:

- 8,500 hours of authentic real-life content for 14 languages from Spanish to Indonesian
- unique professionally crafted courses based on empirically proven didactic strategies and enhanced with technology
- cognitive techniques that move new vocabulary to the learner's long-term memory
- all courses tailored to the user's native language and proficiency

[3] www.Babbel.com

- various topics ranging from basic conversations to travel and business
- audio examples and dialogues recorded with real native speakers
- speech recognition to enhance pronunciation
- offline functionality to use the app even without an internet connection
- international customer service available via phone, chat and email

This study was funded by Babbel but the data collection and the analysis were carried out independently by the Research Team. The language test used in the study was designed and developed by an external independent testing company.

Research Design

The random sample for this study was selected from existing or new Babbel users in the US and Germany. People who lived close to Berlin or New York City (NYC) were asked to take a proctored test in Babbel's offices in Berlin and NYC. The rest of the participants could take the online tests at their home computers. In this report we will generally refer to the "Berlin sample" and "New York sample" but they also have people from other areas in Germany and US respectively. There were some additional requirements for the potential participants. They had to be:

- Willing to study Spanish using only Babbel for two months, and come to the testing location for two sets of language tests;
- At least 18 years of age;
- 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 age limitation was in place because the placement test was designed as college placement test.

The recommended goal for the participants in the study was to use Babbel for at least 16 hours during the two-month study, or two hours per week. Based on our experience with previous studies we imposed an absolute minimum 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.

The Spanish language was selected as one of the most 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 June 2016 and August 2016. People who successfully completed the study were given a lifetime free subscription to the premium edition of Babbel for themselves and one friend of theirs and a gift certificate for Amazon.com ($50 for NY sample and 50 Euro for Berlin sample).

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).

Sample Description

The entire sample selection process is graphically represented in the Appendix, Figure A1.
E-mail messages were sent out to Babbel users 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 957 people viewed the invitation page and of those 853 successfully completed the Entry Survey. This was the initial pool of respondents in the study.

Initial Pool (N=853)

The initial pool of potential participants consisted of people from Berlin (N=342) and New York (N=511). Specifically, in the New York pool 158 people (31%) were from NYC area and the rest were from other US states: AL, AR, AZ, CA, CO, CT, DC, FL, GA, IA, ID, IL, IN, KS, LA, MA, MD, MI, MO, MT, NC, upstate NY, OH, OR, PA, RI, SC, TN, TX, UT, VA, WA, WI and WY.
In the Berlin pool 322 people (94%) were from Berlin area and the rest (N=20) were from other areas in Germany: Bochum, Flein, Frankfurt, Göttingen, Hamburg, Heidelberg, Laucha, Leipzig, Mörfelden-Walldorf, Naumburg/Saale, Salzburg, Stuttgart, and Ulm.

In the initial pool of potential participants about 60% were female. The age varied from 10 to 84 years of age, with a mean age of 43 years. The pool was very well educated with the majority having some college or college/graduate degree.

Table 2A. Education of the Initial Pool (NY)

Category	N	Percent
1. Less than High school	21	4.1
2. HS diploma or equivalent	31	6.1
3. Some college but no degree	133	26.0
4. College graduate, BA or equivalent	159	31.1
5. Graduate degree: MA, PhD or equivalent	167	32.7
Total	**511**	**100**

Table 2B. Education of the Initial Pool (Berlin)

Category (English)	German Category	N	Percent
1. Secondary school (High school) certificate	Realschulabschluss	67	19.6
2. High school certificate	Hauptschulabschluss	10	2.9
3. Vocational school certificate	Fachhochschulreife	23	6.7
4. Level “A” school	Abitur	84	24.6
5. BA or equivalent	Bachelor	28	8.2
6. MA or equivalent	Master/Diplom/Magister	116	33.9
7. PhD and Post-doctoral	Promotion/Habilitation	14	4.1
Total		**342**	**100**

The majority of the people were employed either full time or part-time (56%) and 12% were self-employed and 13% were retired.

Table 3. Employment Status of the Initial Pool by Site

Category	Percent	
	Berlin (N=309*)	**New York (N=501*)**
1. Unemployed	4.5	5.8
2. Student (Full time)	17.2	12.8
3. Employed (Full time)	38.8	48.3
4. Employed (Part time)	11.3	8.2
5. Self-employed	19.4	11.6
6. Retired	8.7	13.4
Total	**100**	**100**

* Some people declined to answer this question.

English was the native language for 92% of the pool in New York and the remainder (8%) included 24 languages: Arabic, Bulgarian, Chavacano[6], Chinese, Croatian, Farsi, Filipino, French, Haitian Creole, German, Hindi, Hungarian, Italian, Japanese, Korean, Malay, Mandarin, Punjabi, Romanian, Russian, Serbian, Spanish[7], Tagalog, and Turkish.

German was the native language for 94% of the pool in Berlin and the remainder (6%) included 9 languages: Bulgarian, Croatian, Farsi, French, Lithuanian, Polish, Russian, Turkish, and Ukrainian.

Almost 85% described themselves as beginner users or never studied Spanish (Berlin 95%, NY 78%). About 22% of the respondents' spouse, partner, or close friends spoke Spanish (Berlin 16%, NY 25%). A very small proportion (7%) of their parents, grandparents, or great-grandparents spoke Spanish (Berlin 2%, NY 11%).

About 87% of the initial pool had studied a foreign language before (mostly at school or college) and over half (58%) knew a foreign language (Berlin 80%, NY 43%).

The primary reason for studying Spanish was personal interest (45%), followed by business or work (17%), travel (17%), school (2%), and other reasons (7%). Nearly 12% got interested in the language because of the Babbel study. For other reasons the respondents mentioned: “all of the above”, “girlfriend/boyfriend/friend speaks Spanish”, “keep my mind busy”, “planning to move to Spain”, “to talk with Spanish family members”, “church work”, etc.

[6] This person was not eligible for the study because the native language is closely related to Spanish.

[7] The Spanish native speaker obviously was not eligible for the study.

Table 4. Reason for Studying Spanish by Site

Category	Percent	
	Berlin (N=342)	New York (N=511)
1. Business/Work	9.4	21.3
2. Travel	26.9	11.0
3. School	2.0	2.3
4. Personal Interest	38.0	49.3
5. Babbel study	18.1	7.4
6. Other	5.6	8.6
Total	**100**	**100**

Most of the people in the NY pool were native or fluent speakers of English. For the Berlin pool the English proficiency varied from basic command to fluent speaker of English.

Pool of Eligible Participants (N=704)

From the Initial Pool (N=853) we excluded the following ineligible participants:

- People who were younger than 18 years of age.
- People with advanced or fluent Spanish.

Altogether 149 people were ineligible for this study and the final pool of eligible participants for sample selection was N=704.

The pool of eligible participants consisted of people from Berlin (N=320) and New York (N=384). Specifically, in the New York pool 121 people (32%) were from NYC area and the rest were from other US states.

In the Berlin pool 301 people (94%) were from Berlin area and the rest (N=19) were from other areas in Germany: Bochum, Flein, Frankfurt, Göttingen, Hamburg, Heidelberg, Laucha, Leipzig, Mörfelden-Walldorf, Naumburg/Saale, Salzburg, Stuttgart, and Ulm.

In the pool of eligible participants about 63% were female. The age varied from 18 to 84 years of age, with a mean age of 44 years. The pool was very well educated: 67% had some college, or college/graduate degree.

The majority of the people were employed either full time or part-time (54%), 5% were unemployed, 13% were students, 16% were self-employed and 12% were retired.
English was the native language for 92% of the pool in New York and German was the native language for 93% of the pool in Berlin.

All people in the eligible pool were either beginner users of Spanish or never studied Spanish. About 20% of the respondents' spouse, partner, or close friends spoke Spanish. A very small proportion (5%) of their parents, grandparents, or great-grandparents spoke Spanish.
About 86% of the eligible pool had studied a foreign language before (mostly at school or college) and over half (57%) knew a foreign language.

The primary reason for studying Spanish was personal interest (44%), followed by travel (19%) and business or work (16%), school (2%), and other reasons (7%). Nearly 13% got interested in the language because of the Babbel study.

Initial Random Sample (N=391)

The people in the initial sample were randomly selected from the pool of eligible participants. In our previous language research studies the initial sample size was about 200 people based on the expected effect size and the dropout rate of about 50%. In this study of Babbel efficacy we increased the sample size to 400 in order to evaluate the difference between proctored and not proctored tests. About half of the sample took proctored exams in Babbel's offices in Berlin and New York and the other half took the tests on their home computers.
The actual size of the initial sample was 391 people (191 in Berlin and 200 in NY). Of them 152 took the proctored test (96 in Berlin and 56 in NY) and 239 people took the unproctored test.

Geographically, in the New York sample 75 people (38%) were from the NYC area and the rest were from other US states and upstate NY.
In the Berlin initial sample 186 people (97%) were from the Berlin area and the rest were from other areas in Germany.

In the initial sample 61% were female (57% in the NY sample and 65% in Berlin). The age varied from 18 to 78 years of age, with a mean age of 41.5 years. The sample was very well educated with more than 80% with some college, or college/graduate degree.

Table 5A. Education of the Initial NY Sample

Category	N	Percent
1. Less than High school	3	1.5
2. HS diploma or equivalent	9	4.6
3. Some college but no degree	43	22.2
4. College graduate, BA or equivalent	67	34.5
5. Graduate degree: MA, PhD or equivalent	72	37.1
Total	**194***	**100**

* Some people declined to answer this question.

Table 5B. Education of the Initial Berlin Sample

Category (English)	German Category	N	Percent
1. Secondary school (High school) certificate	Realschulabschluss	31	16.8
2. High school certificate	Hauptschulabschluss	4	2.2
3. Vocational school certificate	Fachhochschulreife	11	5.9
4. Level “A” school	Abitur	50	27.0
5. BA or equivalent	Bachelor	15	8.1
6. MA or equivalent	Master/Diplom/Magister	70	37.8
7. PhD and Post-doctoral	Promotion/Habilitation	4	2.2
Total		**185***	**100**

* Some people declined to answer this question.

The majority of the people were employed either full time or part-time (51%), 5% were unemployed, 14% were students, 18% were self-employed and 8% were retired.

Table 6. Employment of the Initial Sample by Site

Category	Percent	
	Berlin (N=168*)	New York (N=192*)
1. Unemployed	4.8	4.7
2. Student (Full time)	17.3	10.9
3. Employed (Full time)	39.3	54.2
4. Employed (Part time)	10.7	5.7
5. Self-employed	22.0	14.1
6. Retired	6.0	10.4
Total	**100**	**100**

* Some people declined to answer this question.

English was the native language for 92% of the initial sample in New York and the remainder (8%) included other languages: Arabic, Chinese, French, German, Haitian Creole, Hungarian, Malay, Mandarin, Serbian, Tagalog, and Turkish.

German was the native language for 95% of the sample in Berlin and the remainder (5%) included other languages: Bulgarian, Farsi, Lithuanian, Russian, Turkish, and Ukrainian.

All participants identified themselves as novice users of Spanish or never studied Spanish.

About 18% of the respondents' spouse, partner, or close friends spoke Spanish. A very small proportion (4%) of their parents, grandparents, or great-grandparents spoke Spanish.

Almost everybody (92%) of the initial sample had studied a foreign language before and 64% knew a foreign language (Berlin 83%, NY 46%).

The primary reason for studying Spanish was personal interest (43%), followed by travel (19%) and business or work (16%), school (2%), and other reasons (5%). About 15% got interested in the language because of the Babbel study.

Table 7. Reason for Studying Spanish by Site

Category	Percent	
	Berlin (N=191)	New York (N=200)
1. Business/Work	9.9	22.0
2. Travel	27.2	10.5
3. School	1.6	1.5
4. Personal Interest	39.3	47.0
5. Babbel study	18.3	12.0
6. Other	3.7	7.0
Total	**100**	**100**

Most of the people in the NY sample were native or fluent speakers of English. For the Berlin sample the English proficiency varied from basic command to fluent speaker of English.

Table 8A. Initial Random Sample: Age and Gender Distribution (Berlin, N*=183)

Age	Female (N)	Male (N)	Total (N)	Total (%)
18-20 years old	4	3	7	3.8
21-30 years old	28	23	51	27.9
31-40 years old	23	15	38	20.8
Over 40 years old	63	24	87	47.5
Total	**118**	**65**	**183**	**100.0**

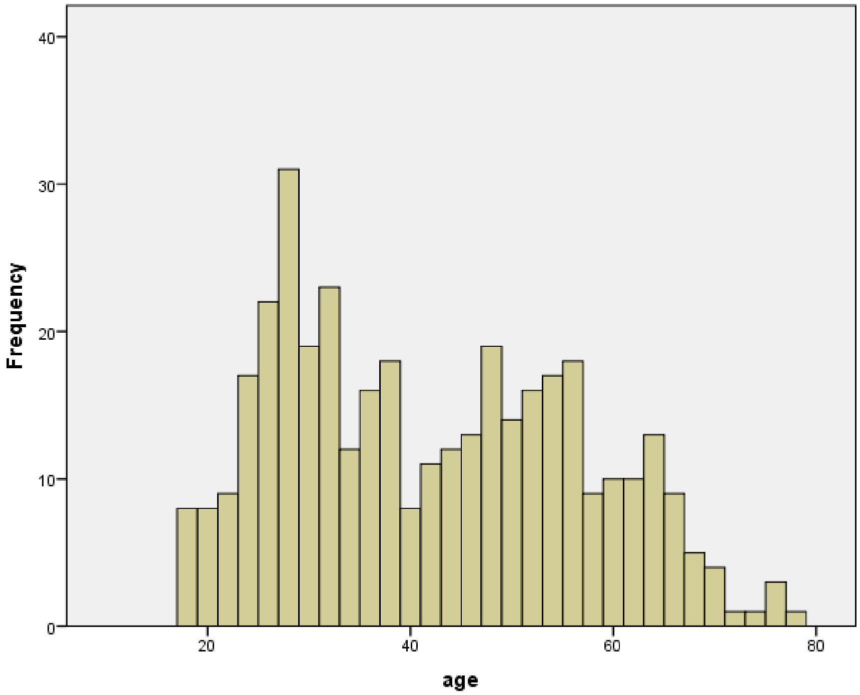
* There are missing data either on age or gender.

Table 8B. Initial Random Sample: Age and Gender Distribution (New York, N*=192)

Age	Female (N)	Male (N)	Total (N)	Total (%)
18-20 years old	5	4	9	4.7
21-30 years old	23	23	46	24.0
31-40 years old	25	14	39	20.3
Over 40 years old	56	42	98	51.0
Total	**109**	**83**	**192**	**100.0**

* There are missing data either on age or gender.

Figure 1. Age Distribution of the Initial Sample.



About half of the sample took proctored exams and the rest used their home computers to take the initial tests. The language proficiency of the initial study sample was as follows:

Figure 2. Initial Test Distribution by Site (WebCAPE points).

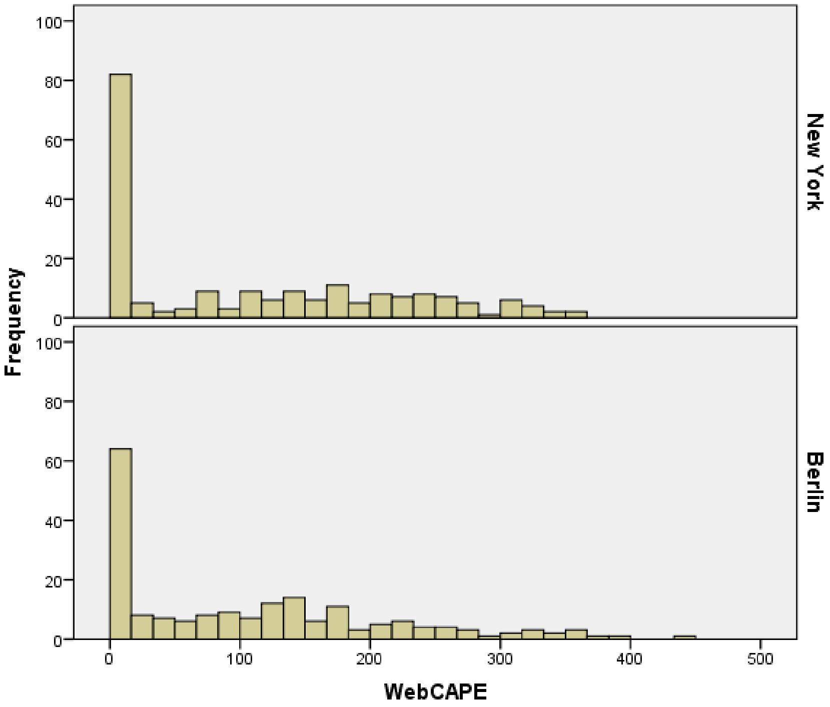


Table 9. Initial Language Proficiency by Site

Statistics	Initial WebCAPE		
	Berlin (N=191)	New York (N=200)	Overall (N=391)
Mean	107.3	105.5	106.4
Standard deviation (std)	106.5	110.8	108.6
Significance (p-value)	Not significant difference (ns)		NA

Table 10. Initial Language Proficiency by Proctored and Not Proctored

Statistics	Initial WebCAPE	
	Not Proctored (N=239)	Proctored (N=152)
Mean	101.8	113.8
Standard deviation (std)	109.6	106.9
Significance (p-value)	Not significant difference (ns)	

Table 11. Initial WebCAPE Semester Placement (N=391)

College Semester	People (N)	Percent
First	356	91.0
Second	24	6.1
Third	10	2.6
Fourth+	1	0.3
Total	**391**	**100.0**

The majority (91%) of the participants were evaluated as novice/beginner users of Spanish and they were placed in First Semester of Spanish. About 6% of the participants were placed in Second semester and 3% in Third or Fourth+ Semester of Spanish. The overall mean WebCAPE score was 106.4 corresponding to First college semester of Spanish.

Final Study Sample (N=325)

The Babbel study continued for approximately two months (8 weeks), starting in June 2016 and ending in August 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 Babbel each week. Based on our previous studies (Vesselinov & Grego, 2012, 2016) the threshold was established at two hours of study as the minimum requirement for the completion of the study.

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 test.
- 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 Babbel to study Spanish for the duration of the study. In the exit survey some 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, YouTube and translation websites and they were allowed to stay in the study. Interestingly enough, people who regularly used additional tools did not have very high results on efficiency. But the number of cases was very small (less than ten) to draw sound conclusions.

The final study sample for language proficiency consisted of 325 people who had used only Babbel, with at least two hours or more of study and valid initial and final WebCAPE tests.

The final study sample (N=325) had a mean age of 41.6 years, from 18 years old to 78 years old, with 62.7% female users. The users were very well educated with 76% having some college or college degree. About 55.8% of them were employed full time or part time, 17.4% were self-employed, 9% were retired, 12.7% were students, and 5% were unemployed.

In the Berlin sample 94% were native speakers of German and the rest included: Bulgarian, Farsi, Lithuanian, Russian, Ukrainian, and Turkish. In NY sample 91% were native speakers of English and the rest included: Arabic, French, German, Haitian Creole, Hungarian, Mandarin, Malay, Serbian, and Tagalog.

About 70% of the sample knew at least one other foreign language (83% in Berlin and 55% in NY) and they were self-reported novice users of Spanish or had never studied Spanish.

About 16% of the respondents’ spouse, partner, or close friends spoke Spanish. About 3% of their parents, grandparents, or great-grandparents spoke Spanish.

Table 12. Final Study Sample: Reason to Study Spanish (N=325)

			Percent
Reason	**Berlin**	**NY**	**Total**
1. Business/Work	9.8	21.7	15.4
2. Travel	26.6	8.6	18.2
3. School	1.2	1.3	1.2
4. Personal Interest	40.5	48.0	44.0
5. Babbel Study	19.1	15.1	17.2
6. Other	2.9	5.3	4.0
Total	**100**	**100**	**100**

The primary reason for studying Spanish was personal interest (44%), followed by travel (18%), business or work (15%), and the Babbel study itself (17%).

Table 13. Final Study Sample: Age and Gender Distribution (N=313*)

Age	Female (N)	Male (N)	Total (N)	Total (%)
18 to 20 years old	8	5	13	4.2
21-30 years old	43	36	79	25.2
31-40 years old	44	24	68	21.7
Over 40 years old	101	52	153	48.9
Total	**196**	**117**	**313**	**100.0**

* There are missing data either on age or gender.

There were 91 (53%) proctored tests in Berlin and 50 (33%) in NY.

People from the final sample used different devices to study Spanish with Babbel and many used more than one device. The majority of them (85.6%) used a desktop/laptop computer; about 25% used a tablet, about 25% used an Android smartphone and 28% used an Apple iOS smartphone. These categories are not mutually exclusive; some people used more than one device. More than half of the participants (52%) used more than one device (39.6% used two devices and 12.4% used three devices).

Final Study Sample vs Not Completed

From the initial random sample (N=391) only 66 (16.9%) 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 test; and people who used additional learning tools during the study. The 16.9% dropout rate is one of the smallest in our language research.

We compared the two groups, the final sample of 325 people and the 66 people who did not complete the study by gender, age, education, employment status, initial knowledge of Spanish (initial WebCAPE score) 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.

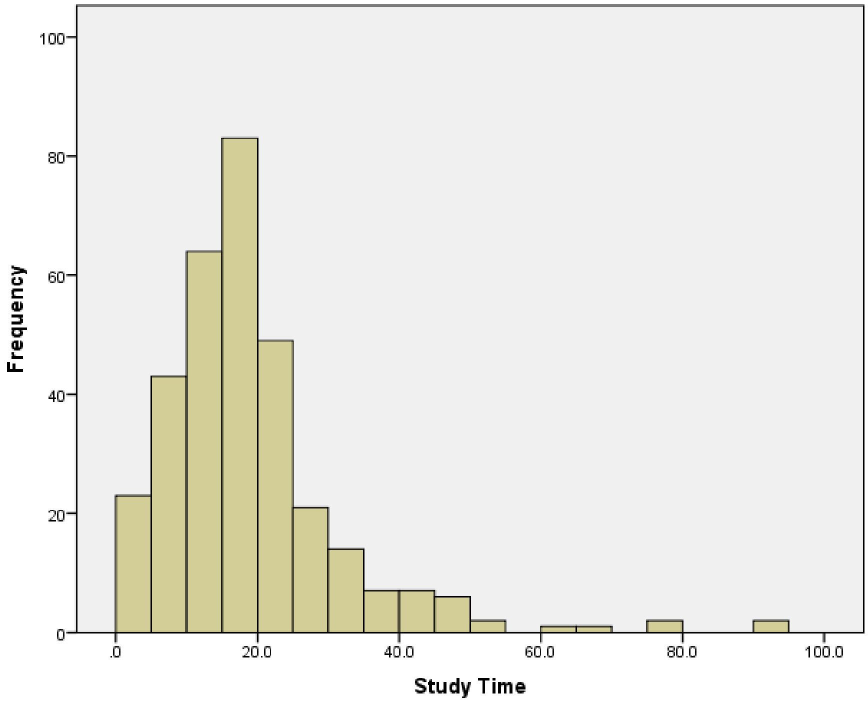
We also compared the sample composition for the two study sites: Berlin and New York. There were no statistically significant differences (at 1% error) between the two sites on age, gender, education (college), and initial knowledge of Spanish (initial WebCAPE score). That is why we have pooled the data from the two sites and reported the results from the combined data.

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=325) was about 19 hours, or a little over two hours a week. The study time varied from about two hours to 93 hours. The average time for the Berlin sample was about 19.7 hours versus 18.2 hours for the NY sample but the difference was not statistically significant.

Figure 3. Study Time (in Hours) Distribution (N=325)



WebCAPE Test Results

All participants took an 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 14. Language Improvement (N=325)

WebCAPE Test Points

Statistics	**Initial WebCAPE**	**Final WebCAPE**	**Improvement (Final-Initial)**
Mean (std)	108.6 (107.8)	265.1 (124.2)	156.5 (121.0)
Median	95	268	147
95% Confidence Interval[8]	96.8 – 120.4	251.5 – 278.7	143.4 – 169.7

The overall average improvement of 156.5 WebCAPE test points was statistically significant with a 95% Confidence Interval from 143 to 170 points. This also means that the improvement

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

in language proficiency for the final sample was statistically significant (at 5% error). Overall 92% of all participants improved their language proficiency (increased their WebCAPE score) with a 95% confidence interval[9] of 88.5% to 94.5%.

Only 26 people out of 325, or 8% did not improve their results. There are two plausible explanations for this fact. First, some of them were more advanced learners of Spanish (second semester or higher) 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.

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 15. WebCAPE Semester Placement (N=325)

College Semester	**Initial Test**		**Final Test**	
	People (N)	%	People (N)	%
First	296	91.1	165	50.8
Second	20	6.2	88	27.0
Third	9	2.8	51	15.7
Fourth+			21	6.5
Total	325	100	325	100

People at First Semester level decreased from 91% to 51% and the proportion of people in Second to Fourth+ Semester level increased notably.

[9] 95% CI with Agresti-Coull correction (Agresti & Coull, 1998).

Main Results

Language Proficiency

Table 16. Language Proficiency Improvement (N=325)

Level (Semester Change)	Improved		Study Time
	People (N)	%	Mean (Hours)
-1 Negative change	4	1.2	15.2
0 Same/No Change	175	53.8	18.4
1 One Semester Up	88	27.1	19.2
2 Two Semesters Up	43	13.2	22.9
3 Three Semesters Up	15	4.6	14.5
Total	**325**	**100**	**19.0**

Overall about 45% of the participants moved up at least one semester. About 27% moved up one semester, 13% moved up two semesters and 5% moved up three semesters. About 55% stayed in the same semester they started in and four people moved down a semester. As the results indicate (last column of the table above), the people who had invested the lowest amount of effort and study time were unsurprisingly the ones who did not improve their language proficiency in semester level. The group of three semesters improvement was too small to draw proper conclusions. People who did not improve their semester standing had studied for 15-18 hours while the rest had on average 19 to 23 hours of study.

The problem with the semester improvement 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 initially has 269 test points (First semester), only a one-point increase in score 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. This measure is not very consistent for interpretation.

The main efficacy measures are presented below.

Table 17. Main Result. Efficacy of Babbel (N=325)

Statistics	**Efficacy Improvement per one hour of study** WebCAPE Test Points	**Time to cover the requirements for one semester of college Spanish** Hours
Mean	12.7	21.2[10]
95% Confidence Interval	10.5 – 15.0[11]	18.0 – 25.7[12]

On average Babbel users will gain 12.7 WebCAPE test points per one hour of study with a 95% confidence interval of 10.5 to 15 test points per one hour of study.

The main measure of Babbel 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, Babbel users will need about 21 hours of study to cover the requirements for one college semester of Spanish with transformed lower and upper limits of 18 hours to about 26 hours of study.

[10] The threshold of 270 points divided by the mean efficacy (12.742 points).

[11] We also bootstrapped (N=10,000) the confidence interval but the result remained practically the same.

[12] The threshold of 270 points divided by the lower limit (10.5) and the upper limit (15.0) of the 95% CI.

Efficacy and the Initial Level of Knowledge of Spanish

Table 18. Efficacy by Initial Semester Level of Language Ability (N=325)

Initial Level	**People**	**Efficacy**
College Semester	N	Mean (std)
First	296	13.5 (21.7)
Second	20	4.6 (7.6)
Third	9	5.0 (8.3)
Total	**325**	**12.7 (21.0)**

The overall efficacy is 12.7 WebCAPE points per one hour of study but novice users of Spanish managed a bigger gain of 13.5 points per hour of study. For the second and third semester levels the improvement was more modest at about 5 points per hour.

Efficacy of Truly Novice Users of Spanish

The larger sample size of this study allowed us for the first time to evaluate the efficacy of truly novice users of Spanish, i.e. users who scored zero on the initial WebCAPE test.

Table 19. Efficacy by Initial Level of Language Ability (N=325)

Initial Level	**People**	**Efficacy**	**95% Confidence Interval**
WebCAPE test	N (%)	Mean (std)	
Truly Novice User (WebCAPE=0)	107 (32.9)	18.1 (28.1)	12.7 – 23.5
Some Knowledge of Spanish (WebCAPE>0)	218 (67.1)	10.1 (15.8)	8.0 – 12.2
Total	**325**	**12.7 (21.0)**	**10.5 – 15.0**

The overall efficacy is 12.7 WebCAPE points per one hour of study but truly novice users of Spanish managed much bigger gain of 18 points per hour of study. Transformed into college level placement this means that truly novice users would need on average about 15 hours[13] of study to cover the requirements for the first college semester of Spanish with range of about 12 to 21 hours[14,15]. of study.

[13] The threshold of 270 points divided by the mean efficacy for truly novice (18 points).

[14] The threshold of 270 points divided by the upper limit (23.5) of the 95% CI.

[15] The threshold of 270 points divided by the lower limit (12.7) of the 95% CI.

Factors for Efficiency

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

None of these potential factors had a statistically significant effect on the efficacy (p=.01). 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 Babbel app works similarly well for people with different gender, age, native language, education, employment status, etc.

The reason for studying Spanish seems to have an effect on the efficacy. People who studied Spanish for business/work tended to have higher efficacy than the rest of the groups (ANOVA with Tukey test for multiple comparisons).

The Babbel study is the first language study where we were able to compare proctored to not proctored test results. The difference between the efficacy of the proctored tests and efficacy of not proctored tests was not statistically significant (p=.01). There were 141 proctored tests and 184 not proctored tests. One interesting observation is that people with proctored tests tend to study a little more than not proctored tests people; on average 21 hours of study vs 18 hours. Also, the proctored test people tend to stay in the study with a dropout rate of only 7%, while for the not proctored test group the dropout rate is 23%. Proctoring the tests does not affect the study results related to the efficacy but it keeps the people in the study and makes them study a little bit more than not-proctored test people. It is our recommendation that it is better for the language studies the tests to be proctored but proctoring is not an absolute requirement.

User Satisfaction

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

Table 20. User Satisfaction (N=313)

Percent

Do you agree with the following statement?	Strongly Disagree/ Disagree	Neither Disagree nor Agree	Agree/ Strongly Agree
"Babbel was easy to use"	3.5	1.6	94.9
"Babbel was helpful in studying Spanish"	1.3	4.2	94.6
"I enjoyed learning Spanish with Babbel"	1.3	6.4	92.3
"I am satisfied with Babbel"	2.6	3.5	93.9

After two months of study, the overwhelming majority of users (92% to 95%) agreed with the positive statements that: Babbel was easy to use, helpful, they enjoyed learning with Babbel and were satisfied with it.

On a different set of questions, participants reported that the lessons helped them improve their ability to communicate in Spanish (81%) and that they were engaged in the lessons (91%).

On the question about how difficult the Babbel lessons were the participants were equally divided: one third thought that they were not very difficult, one third were neutral and the last third thought that lessons were difficult.

Almost all (99%) of the respondents in the exit survey declared that they will continue to use Babbel after the study ends.

In the exit survey a special question was included: "How likely are you to recommend Babbel 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 67% and the "Detractors" were 7.7% and "Passives" were 25.3%. The Babbel NPS was +59.3 which is extraordinary high achievement.

Limitations of the Study

This was the first study for our Research Team in which part of the language tests were proctored and the rest were not proctored by design. We were able to compare the efficacy results for both groups. There were no statistically significant differences on efficacy between the proctored and not proctored tests. The proctored test group had lower dropout rate and higher study time. Proctoring should be preferred but not required.

The WebCAPE test used in this study is not tailored to any specific learning tool, including Babbel. On the one hand, some participants in the study complained that the test sometimes contained words or expressions that were not part of their regular course with Babbel. On the other hand, people insisted that they had learned a lot more than the test asked for. The test is valuable as an independent tool for evaluation which allows us to compare efficacy across different apps, however it does not provide a complete measure of the full progress of users. So their progress evaluation in language proficiency is generally conservative.

There are some limitations of the study, mostly related to the instruments and technological limitations. The online WebCAPE written test measures the progress of beginner/novice users of Spanish well, but it is not very 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 of 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 a little over 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 Babbel. 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.

Conclusion

The Babbel efficacy study is based on a final random sample of 325 people, 18 years of age or older, residing in Germany with proctored sample in Berlin, and US with proctored sample in New York City. All participants were self-reported novice/beginner users of Spanish.

Overall 92% of the participants improved their language proficiency (gained WebCAPE points). The main goal of measuring the efficacy of Babbel was achieved with this study. The results show that, on average, one hour of study with Babbel alone leads to an improvement of 12.7 points on the college placement test WebCAPE. There is a lot of variability of the efficacy and the 95% confidence interval is between 11 and 15 points per hour.

In other words, a Babbel user would need on average 21 hours to complete the requirements for one college semester of Spanish. The transformed upper and lower limits are between 18 and 26 hours of study.

The main factor for the efficacy is the initial level of language knowledge of the participants. The novice/beginner users (First semester) gain faster, with an average of 13.5 points per one hour of study and the more advanced users (Second and Third semester) gain on average 5 points per one hour of study.

The truly novice users of Spanish with initial WebCAPE score of zero, managed much bigger gain of 18 points per hour of study. Transformed into college level placement this means that truly novice users would need on average about 15 hours of study to cover the requirements for the first college semester of Spanish with range of 12 to 21 hours of study.

There are only a handful of known studies with direct objective measures of efficacy of language learning apps. Among them the efficacy of Babbel 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.

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. and J. Grego, 2016, The Busuu Efficacy Study.
http://comparelanguageapps.com/documentation/The_busuu_Study2016.pdf , or
https://blog.busuu.com/wp-content/uploads/2016/05/The_busuu_Study2016.pdf

Vesselinov, R. and J. Grego, 2015, Efficacy of New Language App, report forthcoming.

Vesselinov, R. and J. Grego, 2012, Duolingo Effectiveness Study.
http://comparelanguageapps.com/documentation/DuolingoReport_Final.pdf, or
http://static.duolingo.com/s3/DuolingoReport_Final.pdf

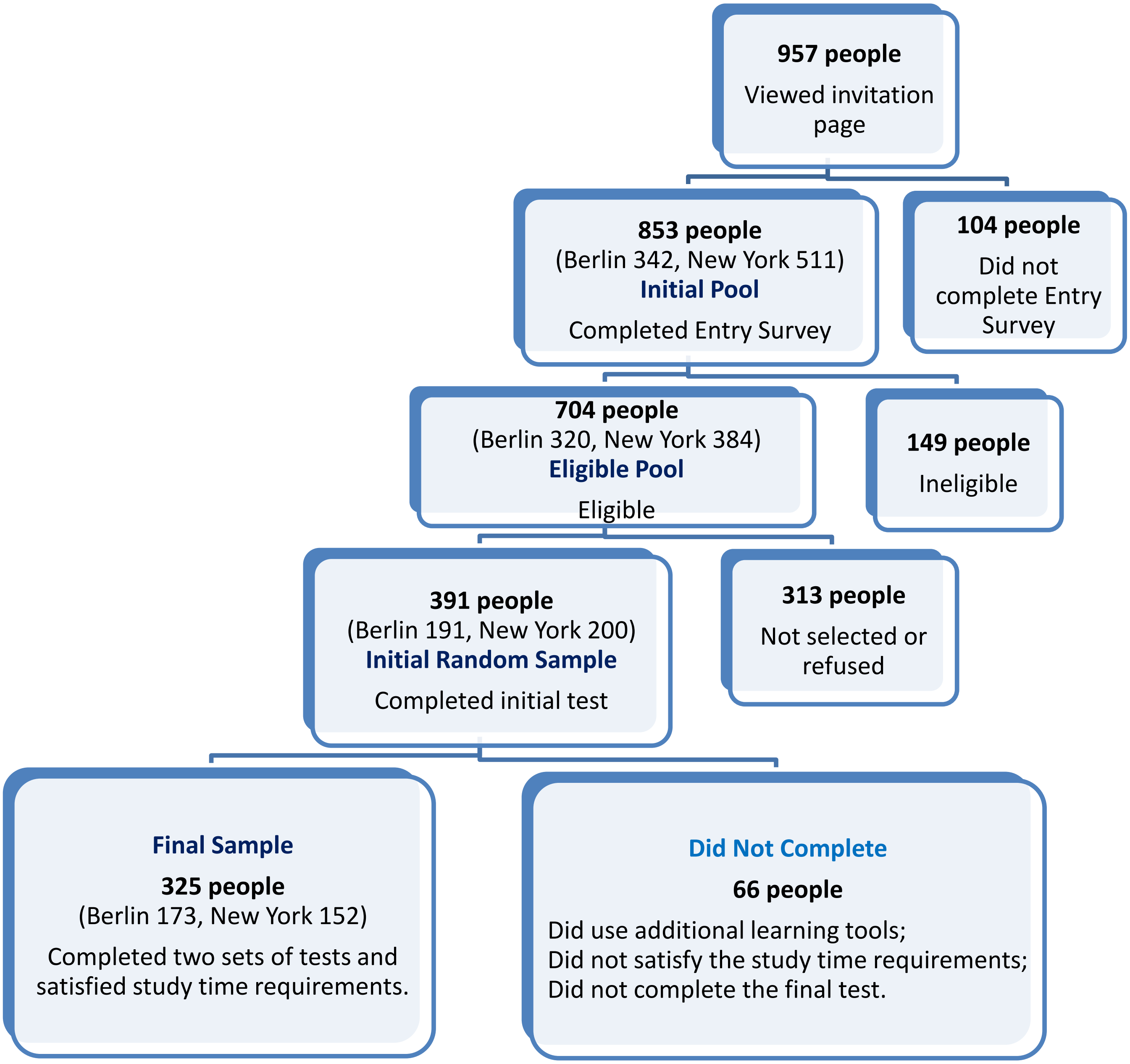
Vesselinov, R., J. Grego, B. Habing, A. Lutz, 2009a, Measuring the Attitude and Motivation of Rosetta Stone® Users.
http://comparelanguageapps.com/documentation/MeasuringTheAttitudeandMotivationofRSUsers.pdf

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

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

Appendix

Figure A1. Sample Selection Tree


957 people
Viewed invitation page
853 people
(Berlin 342, New York 511)
Initial Pool
Completed Entry Survey
104 people
Did not complete Entry Survey
704 people
(Berlin 320, New York 384)
Eligible Pool
Eligible
149 people
Ineligible
391 people
(Berlin 191, New York 200)
Initial Random Sample
Completed initial test
313 people
Not selected or refused
Final Sample
325 people
(Berlin 173, New York 152)
Completed two sets of tests and satisfied study time requirements.
Did Not Complete
66 people
Did use additional learning tools;
Did not satisfy the study time requirements;
Did not complete the final test.