



FINAL REPORT

Roumen Vesselinov, Ph.D.
Visiting Assistant Professor
Queens College | City University of New York
Roumen.Vesselinov@qc.cuny.edu
(718) 997-5444

January 2009

Measuring the Effectiveness of Rosetta Stone®

Executive Summary

MAIN FINDINGS

1. This is the first study that decisively determines the effectiveness of Rosetta Stone Spanish software.

2. After 55 hours of study with Rosetta Stone students will significantly improve their Spanish language skills.

3. After 55 hours of study with Rosetta Stone we can expect with 95% confidence that the average WebCAPE score will be on the level that will be sufficient to cover the requirements for one semester of study in a college that offers six semesters of Spanish.

4. After 55 hours of study with Rosetta Stone Spanish significant proportion of students (56%-72%) will increase their oral proficiency with at least one level.

5. Rosetta Stone Spanish Software is evaluated extremely favorably by its users. After 55 hours of use almost unanimously (80%-90%) the users agreed that Rosetta Stone Spanish software was easy to use, very helpful, enjoyable, and very satisfactory, and that they will recommend this software to others.

PROJECTIONS

6. It is projected, that after 70 hours of study with Rosetta Stone Spanish the majority of

students will complete the requirements for the first semester for any college Spanish course (e.g. 3, 4, 5, or 6 semester Spanish course). This projection needs to be statistically tested further with a follow-up study of 70-100 hours of study with Rosetta Stone Spanish.

7. Rosetta Stone Spanish can be considered equivalent to one semester of six semester course study of Spanish. It is possible to investigate further which of the universities and colleges who use WebCAPE have six semester Spanish course and find the average tuition for one Spanish semester.

The first five findings are backed by statistical sample survey analysis. They are generalizable to the general population. The projections in points 6 and 7 are simple extrapolation of the survey sample results and are not backed by statistical evidence. Additional research is needed to scientifically confirm these two conjectures.

Introduction

This is a research project that focused on studying Spanish as a foreign language. The participants were randomly selected from people responding to advertisement. The respondents were reviewed on demographics

and language skills. People below age of 19 and above 70 were excluded from the pool of potential participants. Also people who declared advanced knowledge of Spanish or, were of Hispanic origin were excluded from the pool. The respondents were asked for their preference for Facility or Remote (at home) participation. This question should have been asked after the sampling was done so their answers were ignored in the selection process. This was necessary in order to avoid the self-selection problem. The participants were randomly selected into two samples: Facility and Remote (At home). If a participant was selected for a sample that s/he did not expressed a desire for, they were dropped from the pool. Again, this was necessary to avoid the self-selection bias. In future research it is recommended that the preference question is not asked beforehand. After the sampling is done, the respondents are approached with the information of their assignment to a particular sample. If they express desire to switch the sample, they will be dropped from the study.

The participants were given equal opportunity to study Spanish in a home environment and at Rosetta Stone Facility. The length of the study was limited to 55 hours which was strictly followed and controlled by the investigators.

Part 1. Sample Description

The participants in the study were randomly selected from a pool of 6409 respondents from the Washington DC Metro Area (Washington DC, Northern Virginia and Maryland). 22% of the pool had some knowledge of Spanish (69% Novice and 31% Beginner). Unfortunately Intermediate, Advanced and Fluent were excluded before frequencies can be produced. 34% of the pool knew another foreign language as well.

The pool was predominantly female (63.3%). Race decomposition was Caucasian (52.6%), Black (34.3%), Asian (8.1%), and Other (5%). The average age was 41 years with the oldest person being 91 years old.

The pool was highly educated with 74.7% having college degree, M.A. or Ph.D. The majority (75.9%) of the pool were employed either FT or PT. The median income was \$50,000-\$75,000, with range from below \$25,000 to more than \$150,000.

The two samples were randomly selected from this pool. Only one person had to switch the group because of extreme circumstances shortly after the beginning of the study.

DEMOGRAPHICS

There were no statistically significant differences between the people who were approached to participate in the study but refused and the people who accepted our offer. We tested for differences on gender, age, race, education, income, foreign language knowledge, and employment status.

210 people were approached for the study. Of them 34 never responded, and 176 started the study and 135 finished successfully. The dropout rate was 23% (41 dropped out of 176). There were no statistically significant differences between the dropouts (n=41) and the final sample (n=135) on gender, age, race, education, income, foreign language knowledge, and employment status. In other words we have no reason to believe that people who dropped out were any different than people who stayed in.

Our final sample consisted of 135 people, divided into two groups: Facility (n=70) and Remote (n=65). There were no statistically significant differences between the two groups on: gender, age, race, education, income, and employment status. The only dif-

ference was on foreign language knowledge with Remote sample having 21.5% with foreign language versus 42.9% for the Facility sample. This parameter was not part of the initial sample design and does not affect our analysis.

The final sample of 135 people had 57% female participants, with mean age of 39 years. Of the whole sample, 78.5% had college degree or above, 80% had full time or part time job and 33% knew foreign language (not Spanish). The median income was between \$50,000 and \$75,000. Below are some detailed distributions by main demographic variables.

Table 1. Race Decomposition

Race	Percent
African American/Black	21.4
Asian	8.4
Caucasian	61.1
Native American/Alaskan	0.8
Other	8.3

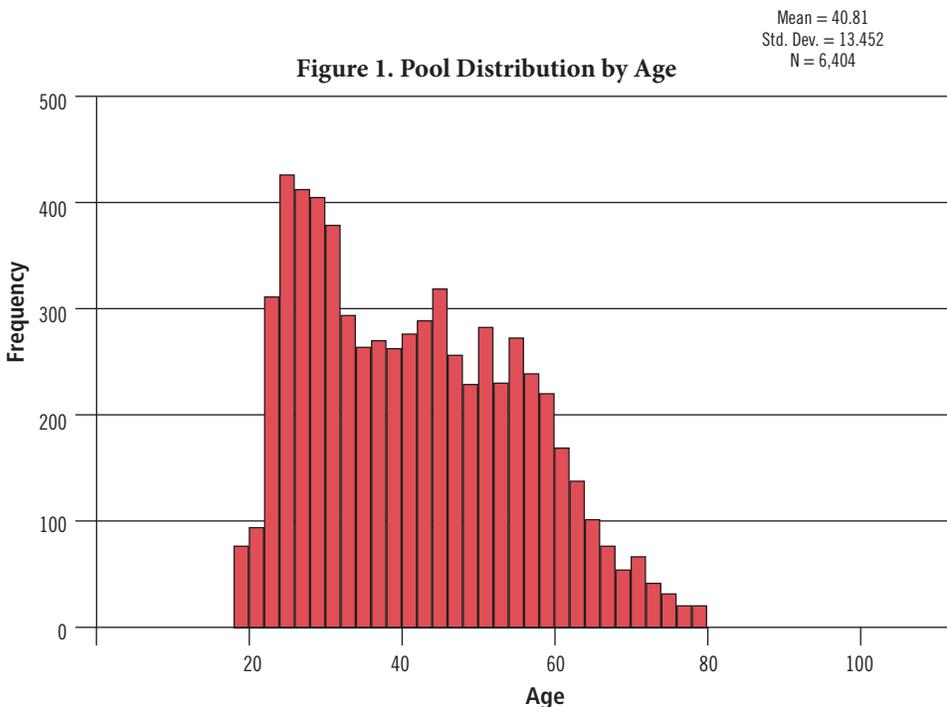
Table 2. Education Decomposition

Education	Percent
High School Diploma/GED	3.0
Some College	18.5
College Degree	42.2
MA/PhD	36.3

Table 3. Income

Income (\$1,000)	Percent
<25	10.0
25-50	25.0
50-75	20.0
75-100	14.2
100-125	11.7
125-150	9.2
>150	10.0

Figure 1. Pool Distribution by Age



Part 2. Software Evaluation

All participants after using Rosetta Stone Spanish software for 55 hours were asked to evaluate 5 statements about their experience with Rosetta Stone Spanish with possible answers ranging from Strongly Disagree to Strongly Agree.

Q1 (Easy). Rosetta Stone Spanish is easy to use.

Q2 (Helpful). Rosetta Stone Spanish is helpful in teaching me the language.

Q3 (Enjoyed). I enjoyed learning Spanish with Rosetta Stone.

Q4 (Satisfied). I am satisfied with Rosetta Stone Spanish.

Q5 (Recommend). I would recommend Rosetta Stones software to others who are interested in learning Spanish.

If we consolidate the “Agree” and “Strongly Agree” we get the results in Table 5.

These results are extraordinary convincing that Rosetta Stone Spanish is extremely easy to use, very helpful, and enjoyable to work with. Finally, between 85% and 91% will recommend this software to others.

Table 4. Software Evaluation

Easy to Use	Easy	Helpful	Enjoyed	Satisfied	Recommend
Strongly Disagree	0.8	0.8	0.8	3.1	3.9
Disagree	1.6	3.1	3.9	5.4	2.3
Neutral	3.1	7.8	7.8	14.7	9.3
Agree	46.5	44.2	40.3	49.6	37.2
Strongly Agree	48.1	44.2	47.3	27.1	47.3

Table 5. Software Evaluation. Confidence Intervals

Dimension	Percent “Agree” or “Strongly Agree”	95% Confidence Interval	
		Lower Limit	Upper Limit
Easy	94.6	90.7	98.5
Helpful	88.4	82.8	94.0
Enjoyed	87.6	81.9	93.3
Satisfied	76.7	69.5	84.0
Recommend	84.5	78.3	90.8

Part 3. Outcome Measures (WebCAPE)

In this study we used as one of our primary outcome measures the WebCAPE test (Web-based Computer Adaptive Placement Exam) developed by the Perpetual Technology Group (<https://www.aetip.com/Products/CAPE/CAPE2.cfm>).

This is a well established¹ test for Spanish with impeccable validity (correlation coefficient = 0.91) and reliability (test-retest = 0.86). According to their website, more than 500 colleges and universities use WebCAPE for placement. Among them are Harvard University, Boston University, Vanderbilt University, Brown University, Queens College, CUNY, University of South Carolina, etc.

The maximum score for Spanish achieved empirically for this adaptive test was 956. The scores are usually a positive number but it is possible to get zero or negative score because of the weights on the

questions. Negative or zero score can be interpreted in a sense that the participant did not take the test seriously or that there were other obstacles because the test is adaptive and every question depends on the answer of the previous question. In that respect negative scores should be interpreted very cautiously or excluded from the analysis.

A student at a college with 6 Spanish courses will need at least 204 points on WebCAPE to move or be placed in Semester 2. Respectively a student at a college with 5 Spanish courses will need at least 234 points; with 4 Spanish courses – at least 270 points, and with 3 courses – at least 281 points.

Table 6. Suggested Calibration Scores

WebCAPE Suggested Calibration Scores							
Spanish: (3) Courses		Spanish: (4) Courses		Spanish: (5) Courses		Spanish: (6) Courses	
Sem 1	Below 280	Sem 1	Below 270	Sem 1	Below 324	Sem 1	Below 204
Sem 2	218 - 351	Sem 2	270 - 345	Sem 2	234 - 311	Sem 2	204 - 288
Sem 3	Above 351	Sem 3	346 - 427	Sem 3	312 - 383	Sem 3	289 - 355
		Sem 4	Above 427	Sem 4	384 - 456	Sem 4	356 - 434
				Sem 5	Above 456	Sem 5	435 - 497
						Sem 6	Above 497

¹ - Personal correspondence with Dr. Jerry Larson, Professor of Spanish Pedagogy, Brigham Young University.

Part 3. Outcome Measures (ACTFL OPIc®)

According to the description on their website (www.actfl.org) “the American Council on the Teaching of Foreign Languages (ACTFL) is the only national organization

dedicated to the improvement and expansion of the teaching and learning of all languages at all levels of instruction. ACTFL is an individual membership organization

of more than 9,000 foreign language educators and administrators from elementary through graduate education, as well as government and industry.”

ACTFL OPIc®

General definition provided by the ACTFL Testing Office

“The ACTFL OPIc® is an internationally used, semi-direct test of spoken proficiency designed to elicit a sample of speech via recorded, computer-adapted voice prompts. Corporations with a need for proficiency evaluations that can be delivered immediately, on-demand will be able to administer an ACTFL Oral Proficiency Interview-like test without the presence of a live tester to conduct the interview.

Completed tests are digitally saved and rated by ACTFL Certified OPIc Raters. The ACTFL Proficiency Guidelines – Speaking (Revised 1999) are the basis for assigning a rating. Research conducted demonstrates that ratings assigned to OPIc samples generally correlate to ratings assigned to direct assessments of speaking proficiency derived through ACTFL Oral Proficiency Interviews (OPI).

The OPIc is intended for all language learners from Novice to Advanced. Large scale testing of spoken language proficiency is now available for secondary and post-secondary language students. The OPIc can be used for placement, formative, and summative assessment purposes. In a business context, the OPIc is appropriate for a variety of purposes: employment selection, placement into training programs, demonstration of an individual’s linguistic progress, and evidence of training effectiveness.

Test Length: Approximately 30 minutes.

Test Format: Digitally recorded prompts are delivered through computer via the internet, or telephonically using VOIP technology.

By Computer: Test is delivered via the internet and taken on computer with a microphone headset. A test candidate moves through the test by “mouse clicking” on navigation aids found on the computer screen. Spoken responses are digitally recorded. At the end of the test, the candidate’s responses are uploaded to the internet for instantaneous delivery to LTI.

By Telephone: Test is delivered by telephone. A test candidate navigates through the test with the aid of verbal instructions and the phone key pad. The candidate’s spoken responses are digitally recorded by LTI.

Test Content: Each test is individualized through the selection of tasks within topic areas according to the test taker’s linguistic ability, work experiences, academic background and interests.

Rating: The OPIc is a criterion-referenced assessment. The ACTFL Certified Rater compares the candidate’s digitally recorded responses to rating criteria as described in the ACTFL Proficiency Guidelines – Speaking (Revised 1999).

Languages: Internet delivered versions of the OPIc are available in English and Spanish.”

Uses of the OPI (not only OPIc): “The ACTFL OPI is currently used worldwide by academic institutions, government agencies, and private corporations for purposes such as: academic placement, student assessment, program evaluation, professional certification, hiring and promotional qualification. The ACTFL OPI is recognized by the American Council on Education (ACE) for the awarding of college credit.

More than 10,000 OPIs in 37 different languages are conducted through the ACTFL Testing Program.”

We used the computerized double rated test (ACTFL OPIc) with 7 levels. These levels are: Novice Low (NL), Novice Mid (NM), Novice High (NH), Intermediate Low (IL), Intermediate Mid (IM), Intermediate High (IH), Advanced (A)

Part 4. Main Results (WebCAPE)

The participants took the WebCAPE test in the beginning of the study (Initial Score) and after completing exactly 55 hours of study with Rosetta Stone Spanish (Final Score).

Figure 2. Initial and Final WebCAPE Score

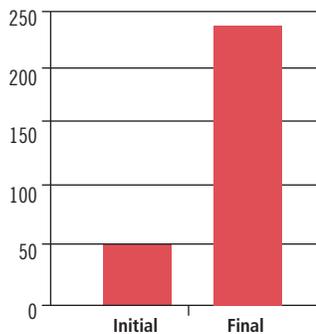


Table 7. WebCAPE Results

Group	Initial Score		Final Score		Change Score = Final - Initial	
	Mean	Median	Mean	Median	Mean	Median
Remote (n=65)	56.2	17.0	235.9	219	179.7	168
Facility (n=70)	49.3	0.0	239.4	243	190.1	177
Total (n=135)	52.6	5.0	237.7	226	185.1	173

There is no statistically significant difference between the Facility and Remote group so the data can be pooled for the analysis.

Table 8. Confidence Intervals (CI) for the WebCAPE Results

Group	Initial Score 95% CI	Final Score 95% CI	Change Score = Final - Initial 95% CI
Remote (n=65)	(38.2 - 74.1)	(210.0 - 261.7)	(153.0 - 206.4)
Facility (n=70)	(32.6 - 66.0)	(214.9 - 263.9)	(161.1 - 219.1)
Total (n=135)	(40.5 - 64.7)	(220.1 - 255.3)	(165.6 - 204.6)

CONCLUSION

1. After 55 hours of study with Rosetta Stone we expect with 95% confidence the average level of WebCAPE score to be between 220 and 255 points.

2. The improvement between the beginning and the end of the study is statistically significant.

PLACEMENT TEST CONSIDERATION

Based on the pooled data study on average one hour of study brings 4.3 points of WebCAPE. The 95% Confidence Interval is (220-255) WebCAPE points. This means that with 55 hours of study with Rosetta Stone 95% of the students will be able to fulfill the requirements for one semester of Spanish with 6 courses. After 55 hours they will reach the placement level for Semester 2 – WebCAPE of 204 points. For the case of 3, 4, and 5 courses this level is not sufficient.

PROJECTION

After 70 hours of study with Rosetta Stone 95% of the students can be placed directly in second semester Spanish courses in any college. This is not a statistically backed conclusion but a linear projection of our results based on 55 hours of study.

Figure 3. Semester Placement by WebCAPE

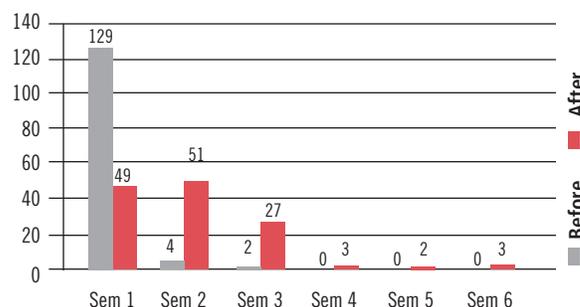
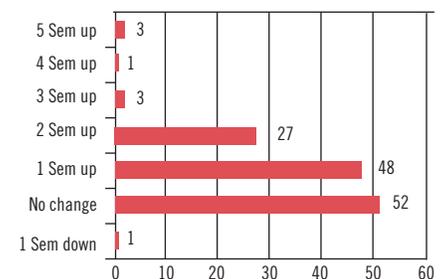


Figure 4. Progress by WebCAPE



If we do a simple projection, based on our study, we can say that after 70 hours study with Rosetta Stone Spanish, 95% of the students could possibly reach at least 280 points of WebCAPE. This score will be sufficient to pass the threshold for one semester for all types of Spanish calibration: with 3, 4, 5 or 6 courses. All of the calculations below are based on the WebCAPE publication “WebCAPE Suggested Calibration Scores.”

Part 4. Main Results (ACTFL)

Table 9. ACTFL Results

ACTFL	Initial		Final	
	Remote % (n)	Facility % (n)	Remote % (n)	Facility % (n)
0 = No Proficiency/Unrated	3.1 (2)	5.7 (4)	1.5 (1)	1.4 (1)
1 = Novice Low	92.3 (60)	87.1 (61)	40.0 (26)	27.1 (19)
2 = Novice Middle	3.1 (2)	5.7 (4)	43.1 (28)	48.6 (34)
3 = Novice High	0 (0)	1.4 (1)	10.8 (7)	15.7 (11)
4 = Intermediate Low	1.5 (1)	0 (0)	1.5 (1)	7.1 (5)
5 = Intermediate Middle	0 (0)	0 (0)	3.1 (2)	0 (0)

The Facility group did a little better than the Remote group but the difference was not statistically significant. The results can be reported and analyzed for the pooled data (n=135).

Table 10. ACTFL Results for Remote and Facility Combined (N=135)

ACTFL	Initial		Final	
	%	(n)	%	(n)
0 = No Proficiency/Unrated	4.4	(6)	1.5	(2)
1 = Novice Low	89.6	(121)	33.3	(45)
2 = Novice Middle	4.4	(6)	45.9	(62)
3 = Novice High	0.7	(1)	13.3	(18)
4 = Intermediate Low	0.7	(1)	4.4	(6)
5 = Intermediate Middle	0	(0)	1.5	(2)

Table 11. ACTFL Results Improvement

ACTFL Improvement	Change Score: Final-Initial %	(n)
0 = No Change/Same*	35.6	(48)
1 = One Level Up	45.2	(61)
2 = Two Levels Up	16.3	(22)
3 = Three Levels Up	2.2	(3)
4 = Four Levels Up	0.7	(1)

The majority of the participants (64.4%) improved their rating on ACTFL.

*One case decreased one level.

Figure 5. ACTFL Improvement

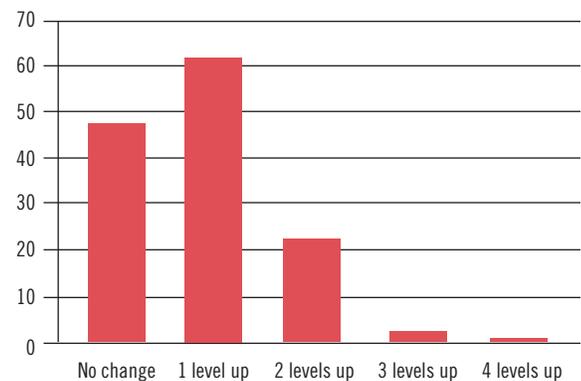
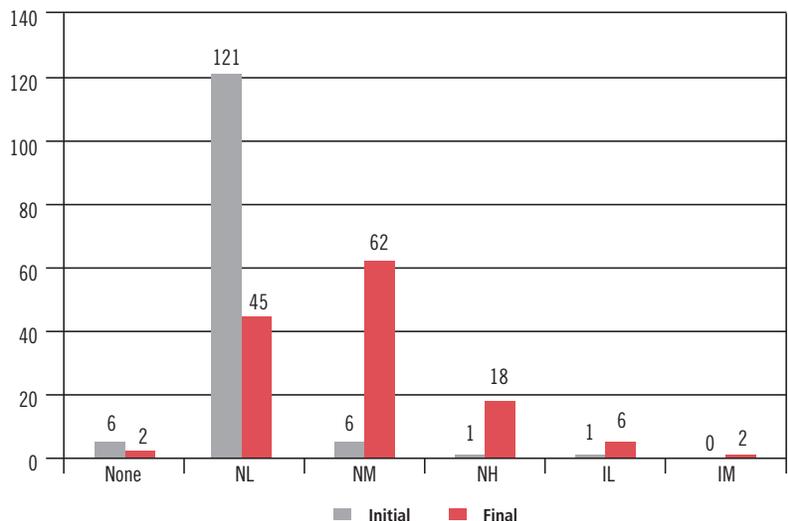


Figure 6. ACTFL Scores Before and After



NL = Novice Low; NM = Novice Mid; NH = Novice High; IL = Intermediate Low; IM = Intermediate Mid

Overall 64.4% of the participants increased their ACTFL results with at least one level. The 95% Confidence Interval for this percentage is (56%-72%). Significant portion of the participants (19.2%) increased their ACTFL with more than one level.

CONCLUSION

After 55 hours of study with Rosetta Stone between 56% and 72% of the students will increase their ACTFL with at least one level.

Part 5. Factors and Outcomes

The purpose of this part of the analysis is to investigate the relationship between the main outcomes and some individual factors and characteristics. We used Chi-square test, Mann-Whitney U test and the Spearman's rho non parametric correlation coefficient.

Foreign language knowledge: People who know foreign language tend to get better scores on the final WebCAPE score ($\rho=.283$) but their improvement on WebCAPE is not statistically different than people who do not know foreign language.

People with foreign language tend to get better scores on the final ACTFL ($\rho=.299$) and they improve their ACTFL more ($\rho=.283$) than people without foreign language.

Gender: No statistically significant influence of gender on the outcomes.

Age: Older participants tend to get better initial WebCAPE scores ($\rho=.186$) but later they tend to get on average lower scores on the final WebCAPE ($\rho=-.215$) and smaller improvement ($\rho=-.278$) than younger people.

Older people tend to get lower scores on the initial and final ACTFL ($\rho=-.210$ and $\rho=-.145$ respectively). But age is not a factor for the improvement on ACTFL test.

Education: Level of education is not a factor for the WebCAPE test and for the initial and final ACTFL. But more educated people tend to get bigger improvement in ACTFL ($\rho=.181$).

Employment: Employment is not a factor for WebCAPE and ACTFL.

Income: People with higher income tend to get lower scores on WebCAPE and have lower improvement than people with lower income ($\rho=-.172$ and $\rho=-.210$). Income has no effect on ACTFL.

Level 1 of Rosetta Stone Spanish: It has no effect on outcomes. The reason is that most people completed this level and there are very few differences here.

Level 2 of Rosetta Stone Spanish: The higher the percentage covered of Level 2 of Rosetta Stone Spanish the better the outcome and the improvement for both WebCAPE and ACTFL. The rho varies from .2 to .4.

Level 3 of Rosetta Stone Spanish: The higher the percentage covered of Level 3 of Rosetta Stone Spanish the better the WebCAPE score and the improvement based on it ($\rho=.374$ and $\rho=.305$). Level 3 has the same effect on ACTFL ($\rho=.274$ and $\rho=.276$).

Effect on Rosetta Stone Spanish Level 1,2,3: People who know foreign language tend on average to cover more of Level 2 and 3 of Rosetta Stone Spanish.

Younger people tend to cover more of Level 1, 2 and 3 of Rosetta Stone Spanish.

Recommendations

This study was one of the first to establish the effectiveness of the Rosetta Stone Spanish software. Its success can be used to further refine this measure of effectiveness. We would like to present the following recommendations for future research.

1. More statistical power is needed in order to satisfy the WebCAPE suggested recommended scores for all types of college Spanish courses (3, 4, 5 and 6 course packages). A new study should require more study hours:

at least 70 and preferably 100 hours for all participants.

2. Very strict control has to be implemented in order to ensure that participants really study, and not just use the software. This control might include periodic WebCAPE tests and other measures.

3. No preliminary self selection of any sort should be allowed.

4. No information collected, particularly for the total pool of respondents should be deleted.

5. Using the Rosetta Stone Facility is preferred but not required. The Facility group did a little better than Remote group but this difference was not significant.