

Vocabulary retention in a spaced repetition longitudinal field study with high-school  
language learners

Miguel A. Varela

Submitted to Swansea University in fulfilment of the requirements for the Degree of  
Doctor of Philosophy

Swansea University

2020

## Abstract

Despite a large amount of research on spaced repetition in L2 courses to retain vocabulary over time, we still do not see its full implementation in everyday classrooms. Laboratory and field studies (on spaced repetition) have worked with participants of different ages and have demonstrated that information can be retained over time, even after several years. Some studies introduced spaced repetition in the classroom, but none of them integrated them fully as part of the curriculum for a whole year.

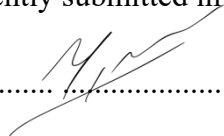
This thesis describes an attempt to integrate spaced repetition in a high-school language course where students take a standard test at the end of the course. To investigate the implementation of spaced repetition, a main research study was conducted in which high-school students rehearsed 100 Spanish words every thirty days in eleven learning sessions. Participants were tested prior and during the treatment to monitor learning. Subjects were also tested 30, 60 and 70 days after the treatment to test vocabulary retention.

Analysis of the results revealed that spaced repetition seems to play an important role in long-term vocabulary retention considering 70 days after the last learning session most of the words were still remembered. Further analysis revealed that the highest retention scores were obtained when the interstudy interval and the retention interval were equal in length. A final important finding was that lack of student motivation and engagement has emerged as a crucial factor that can negatively affect learning and consequent vocabulary retention. The implications of these findings for vocabulary learning research, and for vocabulary teaching in the classroom, are considered.

## Declarations and Statements

### DECLARATION

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

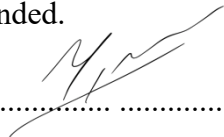
Signed .....  ..... (Miguel A. Varela)

Date ..... January / 28 / 2020.....

### STATEMENT 1

This thesis is the result of my own investigations, except where otherwise stated.


Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed .....  ..... (Miguel A. Varela)

Date ..... January / 28 / 2020.....

### STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organizations.

Signed .....  ..... (Miguel A. Varela)

Date ..... January / 28 / 2020.....

## Table of Contents

Abstract .....	ii
Declarations and Statements .....	iii
Table of Contents.....	iv
Acknowledgements .....	viii
List of tables.....	ix
List of figures.....	xii
Abbreviations.....	xv
Chapter 1: Introduction .....	1
1.1 Relevance of vocabulary in foreign language learning .....	1
1.2 Statement of the problem: Students forgetting vocabulary.....	2
1.3 Repeating .....	4
1.4 Why SR has not been fully implemented in daily L2 classrooms then?.....	5
1.5 Conclusion .....	6
1.5.1 Thesis outline .....	6
Chapter 2: Literature Review .....	8
2.1 RI/ISI Combination to enhance long-term vocabulary retention .....	9
2.1.1 Introduction .....	9
2.1.2 Bahrick (1979).....	10
2.1.3 Cepeda et al. (2008).....	15
2.1.4 Küpper-Tetzel et al. (2014).....	21
2.1.5 Lotfolahi & Salehi (2017) .....	25
2.1.6 Conclusion .....	29
2.2 Teaching and learning.....	30
2.2.1 Word gain vs. retention gain.....	31
2.2.2 Real ecological teaching and learning .....	36
2.2.3 Innovative teaching methods:.....	50
2.2.4 SR to help with authentic materials, motivation, and retention .....	69
2.3 Discussion.....	75



2.3.1	RI/ISI findings.....	75
2.3.2	Learning findings.....	79
2.3.3	Research Questions .....	81
Chapter 3: Replication of Johnson & Heffernan (2006).....		83
3.1	Introduction .....	83
3.2	Summary of the original study .....	84
3.3	Differences between the original vs. the replication study.....	85
3.4	Intervention .....	86
3.4.1	Materials.....	87
3.4.2	Participants .....	90
3.4.3	Method .....	91
3.5	Results .....	93
3.5.1	Vocabulary tests.....	93
3.5.2	Confidence tests .....	95
3.5.3	Retention test.....	97
3.6	Discussion.....	101
3.6.1	Analysis of results from the original vs. the replication study .....	101
3.6.2	Results of the replication study against the literature .....	104
3.6.3	Replication as a pilot study .....	106
3.6.4	Further work and suggestions.....	108
3.7	Conclusion .....	110
Chapter 4: Methodology .....		113
4.1	Introduction .....	113
4.2	Overview of the Study .....	114
4.3	Methodology.....	114
4.3.1	Materials.....	114
4.3.2	Participants:.....	125
4.3.3	Method: .....	129
4.4	Learning and testing sessions.....	140
4.4.1	120-vocabulary test and pre-test .....	141
4.4.2	Session one.....	141
4.4.3	Session two.....	143

4.4.4	Session three.....	145
4.4.5	Session four.....	147
4.4.6	Session five.....	149
4.4.7	Session six.....	150
4.4.8	Session seven.....	150
4.4.9	Session eight.....	152
4.4.10	Session nine.....	154
4.4.11	Session ten.....	154
4.4.12	Session eleven.....	155
4.4.13	Post-test (30-day RI).....	155
4.4.14	Delayed Post-test (60-day RI).....	156
4.4.15	Second delayed post-test (70-day RI).....	156
Chapter 5: Results.....		158
5.1	Introduction.....	158
5.2	Comparison of test scores of the EG vs. the HG.....	158
5.3	Comparison of post-test scores of the EG vs. the CG.....	160
5.4	Experimental group's pre-test, 30-day RI and 60-day RI post-tests.....	164
5.5	70-day RI post-test.....	167
5.6	Results of the tests within the learning sessions.....	170
5.6.1	Tests in session four and ten.....	170
5.6.2	Test in session six.....	172
Chapter 6: Discussion.....		175
6.1	Introduction.....	175
6.2	SR as a means to enhance vocabulary retention.....	176
6.3	Optimal RI/ISI combination.....	182
6.4	More rehearsals may contribute to better learning and retention.....	184
6.5	Vocabulary retention gain.....	186
6.6	Participant motivation.....	190
6.7	Limitations of this study and suggestions for further research.....	192
6.8	Conclusion.....	198

Chapter 7: Conclusion.....	202
7.1 Findings regarding SR and RI and ISI lags.....	204
7.2 Project’s findings regarding teaching and learning.....	205
7.3 Pedagogical implications.....	207
7.4 Final conclusion .....	208
Bibliography .....	210
Appendix I.....	227
Appendix II.....	228
Appendix III.....	230
Appendix IV .....	231
Appendix V.....	233
Appendix VI .....	234
Appendix VII.....	236
Appendix VIII .....	237
Appendix IX.....	239
Appendix X.....	242
Appendix XI.....	244
Appendix XII.....	245
Appendix XIII .....	249
Appendix XIV.....	252
Appendix XV .....	254
Appendix XVII.....	257
Appendix XVIII .....	258
Appendix XIX.....	262

## Acknowledgements

First and foremost, I would like to thank my supervisor Dr. Cornelia Tschichold. I know I could not have done it without the unconditional support and guidance of Professor Tschichold who was extremely patient and encouraging all through the long process of my candidature. To her goes my most heartfelt gratitude.

I would also like to thank my parents and my brother who taught me to always persevere and that with effort anything is possible.

To my wife, Yanina and my son Manuel, I would like to thank for their patience and words of encouragement, giving me time and space when I needed to work.

To Dr. Federico Chachagua I would like to say *¡muchas gracias!* for his words of wisdom and open heart that helped me go through difficult times.

My fellow instructors Imanol and Claudia were extremely supportive and helpful by acting as collaborators in this thesis.

I truly appreciate the willingness and commitment shown by the students at Qatar Academy who participated in the study. Without their support and participation this research project could not have been completed.

I finally would like to acknowledge every single person who in some way or another helped me and encouraged me to finish my thesis.

## List of tables

Table 2-1: Mean percentage of correct recall of Spanish words on first test trials of successive relearning sessions (Bahrick, 1979, p. 300).....	12
Table 2-2: Mean percentage of correct recall of Spanish words on first test trials before and after an increase in the interstudy interval (Bahrick, 1979, p. 301).....	13
Table 2-3: RI groups with their assigned ISI conditions (Cepeda et al., 2008, p. 1097) .....	17
Table 2-4: Estimated ISI days for given RI days (Cepeda et al. 2008, p. 1099) .....	19
Table 2-5: Number of correct responses per test and percentages (Fitzpatrick et al., 2008, p. 241) .....	33
Table 2-6: Tests mean percentages with standard deviation (Bloom & Shuell, 1981, p. 246) .....	39
Table 2-7: Mean percentage of final tests for both conditions, plus SD (Goossens et al., 2012, p. 969) .....	47
Table 2-8: In-class and Out-of-class activities per group (McClean et al, 2013, p. 87).57	
Table 2-9: Vocabulary size mean of pre and post-test scores (McClean et al, 2013, p. 91) .....	58
Table 2-10: Percentage scores comparison of TOEIC and VST tests (Milliner, 2013, p. 56) .....	62
Table 2-11: Mean percentage of correct responses of the three tests (Gryzelius, 2016, p. 16). .....	67
Table 2-12: Mean scores of project tests with standard deviation between parenthesis (Johnson & Heffernan, 2006; p. 73) .....	72

Table 2-13: Comparison of retention gain showing percentage means .....	78
Table 3-1: Replication pre-test/post-test results .....	93
Table 3-2: Test of normality (Shapiro-Wilk).....	94
Table 3-3: Paired samples <i>t</i> -test.....	95
Table 3-4: Descriptive statistics of both confidence tests.....	96
Table 3-5: Test of normality (Shapiro-Wilk) for confidence tests .....	97
Table 3-6: Paired samples <i>t</i> -test for confidence.....	97
Table 3-7: Mean scores of the three vocabulary tests.....	99
Table 3-8: Test of Sphericity .....	99
Table 3-9: Replication ANOVA test.....	100
Table 3-10: Post-hoc comparison of replication tests.....	100
Table 3-11: Word gain comparison among different studies at 0 RI.....	105
Table 4-1: Tests in the study and their purpose .....	136
Table 4-2: Division of Participants by group and length of involvement in the study. .....	137
Table 5-1: EG 30-day RI test vs. HG's test .....	158
Table 5-2: Test of Normality (Shapiro-Wilk).....	159
Table 5-3: Independent samples Students' <i>t</i> -test (EG vs. HG) .....	160
Table 5-4: Pre-test EG vs. CG .....	161
Table 5-5: Test of Normality (Shapiro-Wilk) for pre-tests.....	161
Table 5-6: Independent samples <i>t</i> -test EG vs. CG.....	161

Table 5-7: Test of Normality (Shapiro-Wilk) for post-tests .....	162
Table 5-8: EG 30-day RI test vs. CG's post-test .....	163
Table 5-9: Independent Samples Students' <i>t</i> -Test (EG vs. CG).....	163
Table 5-10: EG's pre-test, 30-day and 60-day RI post-tests .....	165
Table 5-11: Test of Sphericity .....	165
Table 5-12: ANOVA test results of pre-test, 30-day RI and 60-day RI tests .....	166
Table 5-13: Post Hoc Comparisons - Retention .....	166
Table 5-14: EG's pre-test, 30-day, 60-day, and 70-day RI post-tests .....	168
Table 5-15: Test of Sphericity for 70-day RI test .....	168
Table 5-16: Within subjects effects for 70-day RI post-test .....	168
Table 5-17: Post Hoc Comparisons - 70-day RI post-test .....	169
Table 5-18: Results of the tests in session four and ten.....	171
Table 5-19: Test of Normality (Shapiro-Wilk) session 4 and 10.....	171
Table 5-20: Paired samples <i>t</i> -test session 4 and 10.....	171
Table 5-21: Pre-test & session-six test mean values.....	173
Table 5-22: Test of normality (Shapiro-Wilk) test in session six .....	173
Table 5-23: Pre-test & session-six <i>t</i> -test .....	173
Table 6-1: Number of target words (in SR condition) and learning sessions in each study.....	187

## List of figures

Figure 2-1: Recognition test results showing mean accuracy in each RI/ISI combination (Cepeda et al., 2008, p. 1098) .....	18
Figure 2-2: Production test results showing mean accuracy in each RI/ISI combination (Cepeda et al., 2008, p. 1098) .....	18
Figure 2-3: Mean of correctly recalled vocabulary on the final cued recall test as a function of lag and retention interval (Küpper-Tetzel et al., 2014, p. 15) .....	23
Figure 2-4: Project design (Lotfolahi & Salehi, 2017, p. 4) .....	26
Figure 2-5: Mean percentage of correct recall of massed and spaced items (Lotfolahi & Salehi, 2017, p. 11) .....	28
Figure 2-6: Study organization, RI days counted from the last learning session (Fitzpatrick et al., 2008, p. 241) .....	32
Figure 2-7: Memory effect for school A (Erbes et al., 2010, p. 125) .....	53
Figure 2-8: Memory effect for school B (Erbes et al., 2010, p. 126) .....	53
Figure 2-9: Project schedule (Gryzelius, 2016; p. 12) .....	67
Figure 3-1: Screenshot of five questions in one of the webpages of the replication study .....	88
Figure 3-2: Screenshot of a trailer being played with subtitles .....	89
Figure 3-3: Screenshot of a reading with mouseover function activated .....	90
Figure 3-4: Sample question of the replication pre-test .....	91
Figure 3-5: Replication study timeline .....	92
Figure 3-6: Mean difference between performance tests .....	94



Figure 3-7: Mean difference between performance tests.....	96
Figure 3-8: Mean values of replication tests.....	99
Figure 4-1: Sample of the 120-word pre-selection test.....	118
Figure 4-2: Quizlet flashcard sample.....	120
Figure 4-3: Quia test sample.....	122
Figure 4-4: Quia statistical sample.....	123
Figure 4-5: Screenshot of Space Race on Socrative.....	124
Figure 4-6: Project RI schedule.....	132
Figure 4-7: Study organization, RI test days counted from the last learning session	134
Figure 4-8: Screenshot of a pre-test sample question.....	135
Figure 4-9: Project timeline with test and group involvement.....	138
Figure 4-10: Screenshot of the activity to answer in English.....	144
Figure 4-11: Screenshot of a game to practice the target words.....	144
Figure 4-12: Screenshot of the activity to answer in Spanish.....	145
Figure 4-13: Sample <i>meme</i> created by participants.....	147
Figure 5-1: EG 30-day RI test vs. HG's test.....	159
Figure 5-2: Pre-test EG vs. CG.....	162
Figure 5-3: EG 30-day RI test vs. CG's post-test.....	163
Figure 5-4: EG's pre-test, 30-day and 60-day RI post-tests.....	166
Figure 5-5: EG's pre-test, 30-day, 60-day, 70-day RI post-tests.....	170
Figure 5-6: Boxplot of results of the tests in session four and ten.....	172

Figure 5-7: Pre-test & session-six test mean values .....	174
Figure 6-1: Test result comparison of the three groups (EG, CG and HG).....	179
Figure 6-2: Difference between SR vs other conditions across studies.....	181
Figure 6-3: EG's long-term retention results with trendline .....	183
Figure 6-4: EG's project tests.....	185
Figure 6-5: Vocabulary retention gain across studies (30- to 42-day RI).....	188

## Abbreviations

CG	Control Group
EG	Experimental Group
HG	Historical Group
IB	International Baccalaureate
ISI	Interstudy (session) interval
L1	First language
L2	Foreign language
M	Mean
Mdn	Median
MR	Massed repetition
RI	Retention Interval
SD	Standard Deviation
SR	Spaced repetition

## Chapter 1: Introduction

### 1.1 Relevance of vocabulary in foreign language learning

Vocabulary is one of the most important aspects of language learning, and scholars have stressed the fact that without grammar something can still be achieved, but nothing can be achieved without vocabulary (Wilkins, 1972). The magnitude of vocabulary is such that it seems to be that the major difference between native speakers and foreign language learners lies in the size of their mental lexicon itself (Laufer, 1998). Another striking fact showing the importance of vocabulary knowledge is that large amounts of it are supposed to be known to succeed in communication. For example, Laufer (1989) found that 95% of vocabulary needs to be known to comprehend written text successfully, and Hu & Nation (2000) claimed that 98% to 99% of the words in a text are needed to comprehend written discourse. Nation (2006) estimated that for the English language around 8000 to 9000 word families are needed to read authentic materials (e.g., novels, newspapers). The fact that such a large number of words are needed for comprehension highlights the relevance of vocabulary acquisition in foreign language (L2) learning. This fact about vocabulary is particularly overwhelming for beginner language learners aiming to comprehend authentic texts.

The importance of vocabulary acquisition in language learning is also clearly reflected in the structure of some language courses. For example, in high-schools following the IB diploma program (see [www.ibo.org/diploma](http://www.ibo.org/diploma)) students take a foreign language course. Spanish Ab Initio is an example of such a course. The structure of the course itself is generally organized by topics (such as: personal information, family, city life, country life, and health) that rely almost entirely on vocabulary where large amounts of words need to be learned in a rather short period of time. Students in this two-year course are supposed to be complete beginners to be admitted to the course, and they are expected to comprehend rather short and simple authentic texts in the second year of the course. Considering the large amount of words needed to comprehend authentic materials, teaching and learning in such a course could be a daunting experience. Teaching as much vocabulary as possible, which could be a

simple and straightforward strategy in such a course, still does not really seem to work effectively and leads us into the main motivation of this thesis.

## **1.2 Statement of the problem: Students forgetting vocabulary**

The Spanish Ab Initio course is a very particular course considering that no matter how much teaching and learning would normally take place in class, students still tend to forget a large part of what they have learned.

At the school where this research project was conducted, all stakeholders involved (board of directors, administration, teachers, students and parents) were particularly interested in test results since good grades would help with university admissions. Also, a large number of school graduates being admitted to prestigious universities is generally a good promotional asset for any school. This resulted in classes being completely test oriented and students would, for the most part, be focused and willing to succeed. Spanish Ab Initio was no exception, specially, since the students in the program, in their large majority, were willing to pursue their university education in the United States and showed particular interest in learning Spanish as a foreign language.

This particular interest in doing well in tests, seems to be double folded and rather problematic for students as well. Previous research has shown high levels of stress among IB students (Suldo et al., 2008 and 2009) which, in foreign language learning, could negatively affect their overall performance specially vocabulary uptake and long-term vocabulary retention.

Finally, a no-homework policy running at the school prevented teachers from assigning supporting homework activities. Considering the time constraints to bring complete language beginners to a level where they could work with authentic materials, any extra exposure to the language outside of class could have probably contributed to further vocabulary acquisition and retention. Implicit learning activities (e.g., watching movies, listening songs, reading books) could have most probably helped students be implicitly exposed to vocabulary providing more exposure to the words, hence enhancing retention.

The teacher of the course, on the other hand, was confronted with the pressure and the dilemma of having to prepare complete beginners so they can pass their course requirements even when the conditions for doing so are not ideal. Even the implementation of a system that appears to be effective for acquiring large amounts of vocabulary quickly, such as flashcards (Nakata, 2011), could still be problematic. If lessons are too repetitive, the teacher runs the risk of students losing motivation and getting tired of doing the same activities over a two-year period. What seems to complicate matters even more is the fact that, as a rule of thumb, Ab Initio language courses tend to cover the topics mentioned in the previous section (e.g., personal information, family, city and countryside, hobbies) over two academic years. This generally leads to a course structure where a new topic unit is introduced about every four weeks. Although this exposes learners to a wide variety of vocabulary fields, and ideally to a large number of words, it also generates a major problem. It seems then that no matter how much students learn during the first year of the course, they still tend to forget most of what they have learned by the time they reach the second year of the course. Finally, about forty days of study leave before the final exam also negatively affected retention. Students reported that during the time provided by the school to prepare for final exams, they would generally study for one subject at a time a day before the final exam. After taking the exam, they would start to revise for the final exam of the following subject. Spanish Ab Initio always tended to be the last exam students would take, which meant that students would most probably have over thirty days since their last Spanish class without revising. This obviously negatively affected their vocabulary retention.

This notion of forgetting of information has been explored by scholars in the past claiming that (in language courses structured like this) although vocabulary related to a particular topic seems to be properly acquired by the end of every individual unit, this knowledge will not survive substantial periods of time (e.g., Dempster, 1988; Schmidt & Bjork, 1992; Bahrick, 2005). Together with that, in general, in language Ab Initio courses, students tend to be considerably more motivated and committed in their first year of the program, but motivation and commitment diminish the following year. Their learning curve follows the same path. Less student engagement together with lower levels of improvement causes major problems for teachers trying to keep

motivation and student learning at a maximum. Under these circumstances, levels of forgetting tend to increase, hindering also students' success.

Therefore, the main motivation of this thesis was to explore a teaching method that would reinforce vocabulary acquisition while enhancing (large amounts of) vocabulary retention at the same time. Hence considering the context described above (in which in a very short period of time complete beginners need to be able to work with authentic materials, but in which motivation and stress play an important role), there seems to be imperative need of a teaching methodology that would facilitate quick and effective learning (e.g., flashcards), reinforced by a method that has been studied for long and appeared to reinforce learning and improve retention of information such as regular repetition.

### **1.3 Repeating**

There is an apparently simple notion stating that in order to avoid forgetting, information should be repeated. This notion of revisiting or repeating information is based on the claim that what we learn is quickly forgotten immediately after learning (e.g., Anderson & Jordan, 1928; Bahrick, 1979). However, this information can still be kept in memory if revisited periodically.

Although prior to this thesis there used to be a system of repetition through revision lessons in the Spanish Ab Initio course at the school where the main project of this thesis took place, it did not seem to be enough to avoid forgetting. The revision system in place consisted of vocabulary (about 75 words per unit) being revisited in class once a month before tests. There was also a second, more comprehensive revision, that took place just before the end of the first year of the course, and there was a final revision at the end of the second year of the course. For example, for the topic of Free time, after about four weeks of instruction there was a general unit revision at the end of the month before a monthly exam. At the end of the first academic year there was always a week left for revision when all topics covered through the year were revisited. For instance, the same 75 words related to Free time were revised again during that week. Supposing words had not been encountered in any other unit (which was the case with the great majority of the words learned in

every individual unit), words would be usually dealt with for about a month during the main topic in which they appeared, then revised before the monthly test, revisited quickly in a general revision at the end of the first school year, and finally reviewed again at the very end of the course the following year. In this example this generally resulted in good learning during the first month, many of the words would be forgotten by the end of the first year, and the great majority of the words had been forgotten by the time students reviewed vocabulary in class for the very last time in the course, just about a month before their final external examination.

As seen in the previous paragraph, the repetition system in place in that language course was not being effective. The reason seems to be the fact that repetition needs to be structured and needs to follow certain specifications to produce expected results. A long body of research has explored the optimal way of implementing repetition to enhance vocabulary retention and there appears to be robust evidence in the literature to claim that indeed spaced repetition (SR) contributes to long-term vocabulary retention (e.g., Thorndike, 1908; Bahrick & Hall, 1991; Ebbinghaus, 2013).

#### **1.4 Why SR has not been fully implemented in daily L2 classrooms then?**

Despite the fact that spaced repetition seems to play an important role in long-term vocabulary retention it is still not fully implemented in educational institutions. For example, curriculum designers and textbook writers do not seem to embrace vocabulary research and introduce a more systematic approach to vocabulary acquisition, such as recycling vocabulary for instance (Schmitt, 2019). This could partly be attributed to the fact that researchers still seem to disagree when it comes to providing clear guidelines that could be easily transferred to everyday classrooms. For example, even when spacing repetitions at intervals seems to benefit retention of vocabulary, the length of those intervals seems to also determine how long that vocabulary will actually be retained for.

Considering this discrepancy among scholars, it is not intriguing then that spaced repetition is yet to be seen applied in everyday language courses. Some scholars state that more research is needed to finally comprehend how spaced repetition could ultimately be applied to everyday classrooms. For example, Kornell (2009) complains



about the lack of field studies, and there also seems to be a need for more research to see how spacing vocabulary can benefit younger learners (Moss, 1995).

At the teacher level, however, language educators seem to be unaware of the benefits of spaced repetition for long-term vocabulary retention. The researcher interviewed eight fellow language teachers at the school (where this study took place) asking their opinion on spaced repetition. All the teachers interviewed were over forty years of age and had been teaching languages for over fifteen years. All of the teachers acknowledged the benefits of repeating to enhance learning but had not heard about spacing those repetitions over time in a structured manner to avoid forgetting. This also contributes to the answer of why spaced repetition has not yet been implemented in educational settings. Classroom teachers seem to be oblivious of the fact that systematic spaced repetitions can contribute to vocabulary retention.

## **1.5 Conclusion**

The Introduction chapter provided a general background of the main topics involved when aiming at increasing retention of vocabulary through a system of repetitions in a high-school language course. Therefore, the main concern that arises is how to enhance vocabulary acquisition, and more specifically, how to ensure that large amounts of vocabulary remain in the brain for longer periods of time.

The literature review in the following chapter will provide evidence of the validity of the proposed topic of this thesis (to investigate the effectiveness of spaced repetition for retention of vocabulary in a two-year language high-school course) and will highlight the need for such a study to be conducted as a result of the articles reviewed.

### **1.5.1 Thesis outline**

As stated above, considering there are still a few unresolved issues regarding the implementation of spaced repetition in actual educational environments, it seems useful to examine the effectiveness of spaced repetition (SR) for retention of vocabulary in a two-year language high-school course. With that end, this thesis included a replication study (Chapter 3) that studies long-term retention, spaced repetition, authentic materials, teaching techniques, and student motivation. The

replication will serve also as pilot study to identify any potential problems with methodology employed, or possible unexpected inconveniences that could be avoided in the main research project.

The main study (Chapter 4) consists of the main longitudinal research project of this thesis. The project investigated the effectiveness of spaced repetition to enhance vocabulary retention in a high-school language course over a period of thirteen months.

Chapter 4 describes the methodology used in the main project. This chapter provides details regarding the materials employed, participants' demographics, overall planning of the study, and it finishes with a detailed recount of every learning and testing session.

Chapter 5 analyzes the results of the main study. These results might provide relevant data regarding retention of vocabulary, and optimal combination of the retention interval and the interstudy interval.

The Discussion chapter (Chapter 6) summarizes the findings of the two studies in this thesis and discusses the implementation of spaced repetition for maximum vocabulary retention gains. The chapter also presents the limitations of the present thesis and suggests directions for future research.

This thesis ends with Chapter 7 which presents a general conclusion together with pedagogical implications of the findings of this thesis for future improved applications of spaced repetition in actual educational settings.

## Chapter 2: Literature Review

The Introduction chapter briefly referred to the fact that in order to remember information, it must be repeated. If the aim is to remember vocabulary for a long time, then, the visits or repetitions of that vocabulary should be spaced in time. Hence, in order to implement spaced repetition (SR) into a high-school beginners language course, aiming at long-term retention of vocabulary, a main question arises: when is information best supposed to be repeated to ensure long-term retention? The other crucial aspect that stands out is proper learning. Poor learning (i.e. when information has not been encoded correctly and has not been stored properly in the brain yet) will undoubtedly lead to poor retention (Kosslyn & Smith, 2006; Nakata, 2017). In this thesis poor retention will be understood in two ways: vocabulary that is forgotten quickly, or vocabulary that stays for long, but not so much of it is remembered.

Learning and retention appear to be two different topics altogether, the former relating to the period when acquisition and vocabulary rehearsals take place, and the latter referring to the period after rehearsals where only retrievals from memory take place. This literature review, then, has been divided into two main sections. Considering scholars (e.g., Cepeda et al., 2006) suggest planning the general organization of the spacing agenda first before delivering the learning sessions, this chapter will begin reviewing articles that focus specially on the interstudy and retention intervals. Later this chapter will review literature related to teaching strategies employed aiming at enhanced vocabulary acquisition and long-term retention.

Although all of the papers reviewed below touch upon aspects of spaced repetition and/or long-term retention in some way, the main focus of these studies and the methodologies used are very different. The first section reviews four papers that investigate the ideal gap between learning sessions, in relation to the retention interval. Those four papers are, by far, not the only ones referring to the relationship between interstudy interval and retention interval. However, they have been especially included in the literature review because they are representative of two main theories relating to the optimal RI/ISI combination (see 2.1.6 below). Of those four papers there are two (Bahrick, 1979; and Cepeda et al., 2008) that are particularly

salient considering their stance regarding the optimal lag between learning sessions (the interstudy interval) for a given retention interval (RI). The second section reviews nine articles focusing on the kind of teaching techniques researchers have employed in order to investigate retention of vocabulary. This section will also be further subdivided. The first subsection refers to vocabulary gain as a result of spacing the learning sessions as seen in the study by Fitzpatrick et al. (2008). This article brings the two main sections in this literature review (retention and learning) together by emphasizing the need to better comprehend the connection between teaching and learning and long-term retention. The second subsection reviews field studies that investigated retention of vocabulary while aiming at keeping the classroom as ecological as possible. The third subsection, in contrast, covers studies that introduced new learning methods aiming at long-term retention. Although not present in every article, the inclusion of online learning platforms will also be discussed here. Finally, the last article to be reviewed, and that needs special attention, is Johnson & Heffernan (2006). This article is the only one that covers many of the aspects mentioned above (long-term retention, spaced repetition, authentic materials, teaching techniques, and student motivation), highlighting also the importance of preparing foreign language learners to deal with authentic materials. This is of particular importance to IB Ab Initio students who, with a limited knowledge of the language, have to tackle authentic materials in final course assignments.

## **2.1 RI/ISI Combination to enhance long-term vocabulary retention**

### **2.1.1 Introduction**

The articles reviewed in this section focus on long-term retention and discuss the optimal combination between the retention interval (RI) and the lag that should exist between learning sessions called interstudy interval (ISI). Three of the articles (Bahrlick, 1979; Küpper-Tetzel et al., 2014; Lotfolahi & Salehi, 2017) investigate L2 vocabulary retention. Although Cepeda et al. (2008) works with trivia facts, rather than vocabulary per se, it is a very distinctive study considering the researchers conducted thousands of tests to examine the optimal combination between the retention interval and the interstudy interval (RI/ISI). These four articles follow different methodologies, employ different teaching and learning techniques, and are

representative of the discrepancies that exist among scholars regarding the ideal RI/ISI combination.

### **2.1.2 Bahrick (1979)**

The first article in this section is Bahrick (1979) which presented some research on the benefits of spaced repetition for retention of knowledge. Findings of this study, in both the learning phase and the combination between the retention interval and the interstudy interval (RI/ISI) were revealing at the time, and still prevail until present times. Bahrick (1979) first focused on the dynamic process involved in acquisition and maintenance of information in a series of laboratory studies. According to the researcher, to maintain knowledge, information must be revisited periodically, but there are losses of information in intervals between each exposure, something prior memory research had failed to explain. In the study the author emphasized the need for spaced repetition both during encoding and consolidation, i.e. during acquisition to properly learn information, and during the consolidation period, in order to keep that information permanently. In terms of repeating lags, the author realized that the RI/ISI combination plays an important role in the retention of the information.

Bahrick (1979) mentions two methods for retention of knowledge. The first method is called Cross-Sectional Adjustment and, according to the author, its goal is to analyze retention of knowledge over extended periods of time, with retention levels persisting over several decades. A real-life example of this method would be to try to estimate how much learners can remember after studying French for a year, if they are tested ten years after the course had finished.

The second method, and the one Bahrick was particularly interested in, is called Successive Relearning and it offers the possibility of controlling the conditions for learning, re-learning, and testing more easily. This second method is described in more detail below since the methodology employed offers a clear insight of how learning and retention levels are affected by the changing duration of the retention interval and the interstudy interval. Bahrick decided to explore the conditions under which information is kept over long periods of time, claiming that the number of rehearsals is not enough to account for loss or retention of information, but that there

should be deeper reasons that explain the whole process better. For example, Bahrick claimed that for information to remain permanently in a person's brain, it should follow a cycle with acquisition, loss, and relearning of information. This means that information is learned at first encounter, but some is lost, so there is a need for further relearning or reacquisition intervals for final acquisition. Later on, information will need to be accessed periodically to guarantee that learned items are retained and not forgotten.

### **2.1.2.1 Summary**

Bahrick conducted two laboratory experiments using the Successive Relearning method. In the first experiment 50 English-Spanish word pairs were re-learned at intervals that varied from a few seconds to 30 days in length trying to determine the effects of the time interval separating relearning sessions.

Participants were 30 undergraduate university students with no previous Spanish knowledge. Ten participants were arbitrarily assigned to each one of the three different groups. The first group had a learning session once daily, the second group had a learning session every seven days, and the third group had a learning session every 30 days.

In this experiment Bahrick introduced a technique called dropout (every time an item was recalled correctly, it was removed from the list containing all the target vocabulary items). Each group went through six sessions. The first session began with a presentation session of all the target words and it ended with a test. Participants were presented with word pairs (English-Spanish) visually for five seconds in a random sequence. In addition to that, the researcher pronounced each Spanish word. This was immediately followed by a test trial in which subjects were presented with an English word, in a random order, and they had to say the Spanish translation for it. The researcher used the dropout technique in which only failed items were used in the following presentation, and so on until all items were recalled correctly. Each one of the following five relearning sessions began with a similar test to the one in the presentation session. This was done to check how much participants could remember at that stage, followed by the dropout procedure.

Interstudy Interval (in days)	Session				
	2	3	4	5	6
1	53%	86%	94%	96%	98%
7	39%	68%	83%	89%	94%
30	21%	51%	72%	79%	82%

**Table 2-1: Mean percentage of correct recall of Spanish words on first test trials of successive relearning sessions (Bahrick, 1979, p. 300)**

Table 2-1 above shows the average number of words remembered of each of the translation tests taken at the beginning of each relearning session. Results revealed that although participants in group three (under the 30-day interval, once a month over six months) group learned very well, the other two conditions showed greater gain. Clearly group one (with a learning session once a day for six consecutive days) and group two (7-day ISI lag) showed better accumulative learning after six relearning sessions. The author also realized that although there was a difference in learning, by the sixth session the 30-day ISI group was scoring almost as high as the other two groups. Therefore, in order to investigate the spacing effect even further, the researcher decided to test retention, and introduced a 30-day interval after the last learning session when items would not be revised.

Following the same learning procedure and using the same participants as before (the researcher added 30 more participants for this experiment), Bahrick decided to conduct another experiment to test the effects of increasing test interval. Therefore, the author divided participants in two main groups, and three subgroups inside each. The first main group had three subgroups of participants that were trained with 0, 1, and 30-day (ISI) interval between sessions respectively and tested 30 days (RI) after the third session. The second main group was also divided in the same way (0, 1, and 30-day lag between sessions) for six sessions, and all subgroups were tested 30 days after the last (sixth) session. Data (see Table 2-2 showing comparative data of the two main groups) revealed that every subgroup with the smaller (ISI) lag (0 and 1 day)

showed faster cumulative learning at short intervals, but the introduction of the lengthened interval (30-day RI) resulted in a decrease in performance. The 30-day ISI (rehearsing every 30 days) group showed the highest improved performance towards the last learning session, and it also obtained the highest mean scores in the final retention test. The conclusive finding in this experiment was that the group with the largest interstudy interval (ISI) took longer to learn most of the information, but when the 30-RI test was taken, this group showed the highest retention gain.

Interstudy Interval (days)	Session					Following the 30-day interval
	2	3	4	5	6	
After three training sessions						
0	77%	89%				33%
1	60%	87%				64%
30	21%	51%				72%
After six training sessions						
0	82%	92%	96%	96%	98%	68%
1	53%	86%	94%	96%	98%	86%
30	21%	51%	72%	79%	82%	95%

**Table 2-2: Mean percentage of correct recall of Spanish words on first test trials before and after an increase in the interstudy interval (Bahrick, 1979, p. 301)**

In order to rule out the possibility of participants rehearsing Spanish words during non-session times, Bahrick conducted another experiment. Using the same procedure of learning sessions and tests as in the second main group above (i.e., 0,1, and 30-day ISI, six learning sessions, and a 30-day RI), the researcher trained a new cohort of 30 participants using 21 Chinese characters. Results revealed a very similar trend in the relearning sessions (0, and 1-day ISI showed higher cumulative learning than 30-day ISI), and retention test results were comparable to those of the initial study. Mean



values for the 30-day RI retention test for the three different conditions were: 0-day ISI = 55, 1-day ISI = 75, and 30-day ISI = 70. Although the 30-day ISI group did not obtain the highest mean value this time, a similar conclusion could be made. The researcher explained that when the interstudy interval is smaller (rehearsals happen more often) cumulative learning is higher, but information is lost at a considerably faster rate (than with larger interstudy intervals where rehearsals are wider apart) if that information is tested in a long-term retention test. At the same time, Bahrick also stated that information can be retained longer if the value of the interstudy interval (ISI) is not significant smaller than the value of the retention interval (RI), and if the value of the interstudy interval is not higher than the value of the retention interval.

Bahrick (1979) concluded stating that if information is intended to be retrieved once a month, an ideal interval for training sessions is one month ( $RI = ISI$ ). The author continued to explain that if the final retention interval is larger than interstudy interval ( $RI > ISI$ ), then information might need to be re-learned.

#### **2.1.2.2 Commentary**

Although dated, this article shows very important facts about learning and retention. Bahrick (1979) highlights the fact that the learning process is an action in progress that needs to be worked on progressively. At first information is learned, but some is lost, so it must be relearned periodically several times to guarantee that information stays in the brain. Once information is finally stored in the brain, it must have periodical visits for it to be maintained and not forgotten. A revealing finding in this article is that even when spaced (as opposed to massed), acquisition is faster at short interstudy intervals (short ISIs), but it is slower at longer lags (large ISIs). As shown in the last two experiments described above, those longer lags showed lower levels of mastery at the beginning of the learning phase, but by the last learning session, learning levels were not so dramatically different from those of shorter lags. The actual benefits of longer interstudy intervals came to light with the introduction of a lapse of no learning between the last learning session and a retention test (RI). As experiment two and three demonstrate, learning with larger interstudy intervals appears to show a slow increase of vocabulary mastery in the learning phase, but that mastery continues to increase (at least in this study) even after learning has stopped.

Findings from this article provide a first notion of the possible interaction between learning and retention of information and it has big implications for language learning. It is clear that Bahrick was more concerned about the interstudy intervals (ISI) rather than different retention intervals (RI), but the study could have provided more decisive knowledge regarding the RI/ISI combination by adding a few more retention intervals of different lengths. Also, in order to further analyze participant responses and performance, the article seems to leave out important information. Bahrick (1979) does not mention the parameter used to define whether participants pronounced words properly or not, and degrees of errors accepted to still consider a word as correct.

To conclude, Bahrick (1979) addresses the question of when exactly to repeat by highlighting the importance of repeating from the moment information is first acquired. The article highlights the notion that information must be learned first, and then it should be maintained through periodic visits to reach long-term retention. This, therefore, defines long-term retention as resulting from three phases: learning, maintenance, and retention. Finally, Bahrick (1979) concludes that for long-term retention the relearning session lags (ISI), should be equally large as the retention interval (RI). This was later challenged by studies that are discussed next.

### **2.1.3 Cepeda et al. (2008)**

In their article: ‘Spacing Effects in Learning. A Temporal Ridgeline of Optimal Retention’, Cepeda, Vul, Rohrer, Wixted, and Pashler were curious to find out how the timing of study events affected retention. The authors noticed that several prior studies (e.g., Bahrick, 1979 and Bloom & Shuell, 1981) had obtained positive results, in different environments (lab and field), when testing spaced repetition (SR) for long-term retention. Cepeda et al. (2008) however, found that there were still several features of spaced repetition that needed to be resolved. The scholars concentrated specially onto how study-time lags can affect final retention. Based on reviews of prior SR studies (e.g., Cepeda et al., 2006), the scholars realized that there was a strong relationship between the interstudy interval (ISI) and the retention interval (i.e., the study time lags, and the lapse between the last learning session and the time when information is finally retrieved). They also found that most spaced repetition studies

used very short interstudy intervals and/or very short retention intervals (RI), and only a few of those studies had retention intervals extending for more than a week. Based on those prior findings the authors noticed that there was a need to conduct longitudinal studies to determine the effects of the relationship between the interstudy interval and the retention interval (RI/ISI) in long-term retention.

As a consequence, the researchers decided to run thousands of training sessions and tests in order to study spacing effects at different ISI lags. As opposed to Bahrick (1979), the researchers were not interested in how long information remained in the brain, but rather in finding the optimal RI/ISI combination which would determine the best time to repeat information (given a certain retention interval).

### **2.1.3.1 Summary**

In their long study, for learning and testing, the authors set up a purposely built website with 32 obscure trivia facts (e.g., ‘What European nation consumes the spiciest Mexican food?’ Answer: ‘Norway’).

The researchers used 1,354 subjects from their own internet memory research database created to conduct long spaced repetition research. Participants were of various ages (mean age was 34 years) and were located in different countries around the world. Subjects were rewarded for participating in the study, and they were entered in a raffle for cash prizes.

The authors created 26 different conditions to test different combinations of interstudy intervals (ISI) and retention intervals (RI), and participants were randomly assigned to those conditions (see Table 2-3 for a list of RIs and corresponding ISIs). Each condition had two learning sessions, separated by different interstudy intervals, and ended with a final exam at a certain preset retention interval. In the first learning session participants had to learn 32 trivia facts and restudied them in the second session. After the prescribed retention interval subjects took a productive test (providing one-word answers to short questions) and a recognition test (working in a multiple-choice activity).

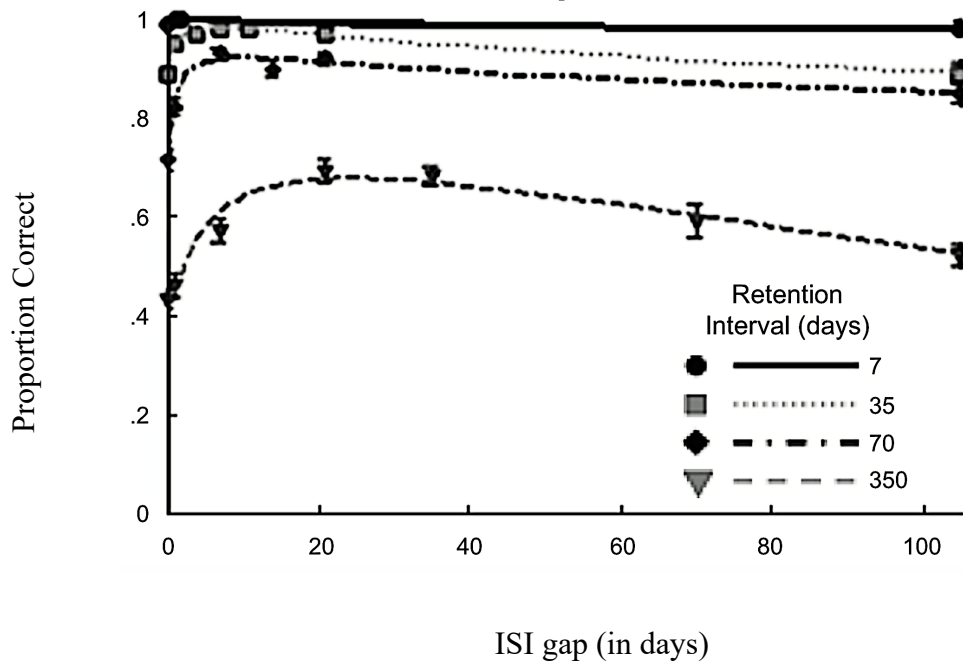
<b>RI Groups (days)</b>	<b>ISI (in days)</b>						
7	0	1	2	7	21	105	
35	0	1	4	7	11	21	105
70	0	1	7	14	21	105	
350	0	1	7	21	35	70	105

**Table 2-3: RI groups with their assigned ISI conditions (Cepeda et al., 2008, p. 1097)**

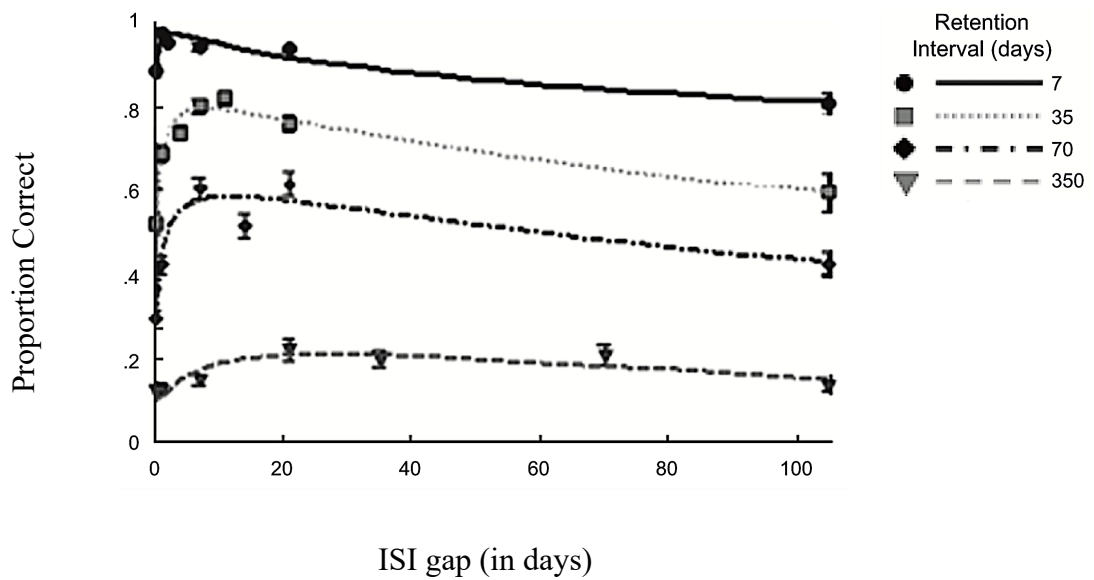
In the first session participants went over each of the 32 facts (presented as questions) and they had to provide an answer to each question. If the answer was correct, then that fact was removed from the list. If the answer was wrong, the correct answer was provided by the system and the question went back to the starting list. Participants had to go over all of the remaining questions, until they had answered all of the questions correctly.

Final testing also took place online, and participants were tested at different time intervals according to the different RI conditions. Subjects received two tests (a productive and a recognition test) in each testing session and feedback was not provided this time.

Overall test results revealed that for each retention interval the first interstudy intervals showed lower overall mean scores increasing as the retention interval increased, but later starting to decrease again (see Figure 2-1, and Figure 2-2 below). The plotted points in the graphics show the mean scores for each test in each RI condition. For example, in the 350-day RI condition, the higher scores were obtained in the 21-day-ISI condition. This means that in order to remember information 350 days after the last learning session, it seems to be best to repeat information about every 21 days (notice that the researchers later interpolated raw data with cubic splines to obtain more precise results. See Table 2-4 for more precise RI/ISI combinations).



**Figure 2-1: Recognition test results showing mean accuracy in each RI/ISI combination (Cepeda et al., 2008, p. 1098)**



**Figure 2-2: Production test results showing mean accuracy in each RI/ISI combination (Cepeda et al., 2008, p. 1098)**

The authors further analyzed data and provided a more precise estimated ideal interstudy interval gap for each RI condition tested (see Table 2-4 below). For instance, in productive knowledge, given a 7-day RI, for higher possible scores, information should be repeated about every three days, which is the 43% of the retention interval.

RI (in days)	Productive knowledge		Receptive knowledge	
	Ideal ISI (days)	Ideal % of RI	Ideal ISI (days)	Ideal % of RI
7	3	43%	1.6	24%
35	8	23%	7	19%
70	12	17%	10	14%
350	27	8%	25	7%

**Table 2-4: Estimated ISI days for given RI days (Cepeda et al. 2008, p. 1099)**

Cepeda et al. (2008) finally reported that study results were conclusive. The researchers stated that typical school courses generally cover a certain topic in a week or couple of weeks, with information (relevant to that same topic) being reviewed during that same week or couple of weeks. This would give a certain feeling of immediate mastery, but information will not be remembered in the long run. The authors asserted that for information to be retained in time, it must be repeated at certain carefully planned intervals. The article finally found that although there is no absolute value for the ideal interstudy interval, it still depends heavily on the retention interval, and it can have a powerful effect on retention if applied properly. Cepeda et al. (2008) concluded that the RI/ISI combination follow an inverted-U shape, having the highest peak at the optimal interstudy interval for a given retention interval, but scores will be lower before or after it.

### 2.1.3.2 Commentary

In the search for an ideal combination of the retention interval and the interstudy interval (RI/ISI) that could produce the best retention levels, this article is particularly important considering the authors used a large sample to produce large amounts of data regarding the ISI and RI relationship. The authors found that the ideal RI/ISI combination is crucial to guarantee long-term retention, that the optimal interstudy interval increases as the retention interval increases, and that the RI/ISI combination follows an inverted-U shape. These three concepts fully agree with Bahrck (1979).

What differs significantly from Bahrck (1979), however, is the fact that, at least for the elicitation of productive knowledge, Cepeda et al. (2008) stated that for a retention interval of 35 days the ideal interstudy interval is around 8 days. Bahrck (1979) had suggested 30-day ISI for a 30-day RI. Cepeda et al.'s (2008) findings are based on extensive reviews of the literature (e.g., Cepeda et al., 2006) and also on a very large sample in their own study. However (as opposed to Bahrck, 1979), since Cepeda et al. (2008) was not related to language learning, it was carried out over the internet, and over two learning sessions alone, it raises the question of whether vocabulary acquisition works in a different way, or whether the methodology employed could have contributed to such different results. At the same time, another important observation arising from Cepeda et al. (2008) is that although participants managed to still remember information in the 350-day RI condition, the amount of information that is remembered differs greatly from any other condition (at round 21% it seems to be very low). Therefore, this article shows that there are still several issues to be resolved in relation to RI/ISI ideal combination.

To conclude, both articles reviewed in this section so far agreed on the fact that there is an optimal interstudy interval for a given retention interval, and that when the retention interval increases, so does the interval between learning sessions. The article reviewed next (Küpper-Tetzel et al., 2014) will also deal with the RI/ISI combination. This article will refer strictly to L2 learning in a real secondary school environment considering that there are still many other variables that need to be tested to see how spaced repetition and long-term retention actually work.

#### **2.1.4 Küpper-Tetzel et al. (2014)**

Küpper-Tetzel, Erdfelder and Dickhäuser were curious about the exact time to repeat in order to enhance long-term retention. As opposed to Cepeda et al. (2008) and Bahrick (1979), the study introduced in this section took place in a real classroom. Just like other articles (e.g., Bloom & Shuell, 1981; Sobel et al., 2011; and Goossens et al., 2012) Küpper-Tetzel et al. (2014) was a field study investigating whether laboratory findings would also hold for secondary school L2 vocabulary learning.

The authors saw that previous investigations had demonstrated that optimal lags between learning sessions increased long-term memory performance in comparison to non-optimal lags (e.g., Küpper-Tetzel & Erdfelder, 2012). At the same time, the authors also found that in previous research (e.g., Bahrick et al., 1993 and Bahrick & Hall, 2005) long-term memory was increased when the interstudy intervals were separated by long lags instead of short lags. Although revealing, this still did not answer the question of when to repeat exactly. The researchers also saw that in Cepeda et al. (2008) for example, although there is no exact lag, the optimal time for repeating (ISI) depended on the length of the retention interval (RI). In the same article the authors also found that memory performance followed an inverted-U-shaped curve. The retention curve (as reflected on results of retention tests) would firstly increase along with the length of interstudy intervals until reaching an optimal (RI/ISI combination) lag and then decrease again. This showed that an ISI lag that is too short or too long could be detrimental for long-term retention.

Considering that the exact combination between the retention interval and the interstudy interval was still not defined, Küpper-Tetzel et al. (2014) studied the effect of different interstudy intervals across different retention intervals. Therefore, the article presented a field study using young learners to investigate how L2 vocabulary could be retained when different interstudy intervals were tested.

##### **2.1.4.1 Summary**

The target vocabulary items for this study were 26 German-English noun word pairs (taken from advanced chapters of the course textbook) that students had not cover in



regular classes yet. The participants were 65 11- to 13-year-old students in a secondary school in Germany from three different classes taking L2 English.

The study was divided into three conditions each one having a different ISI lag (0-day, 1-day, or 10-day lag). Since researchers had to respect classroom structure as the project took place during regular English lessons, an intact cohort method was adopted. Therefore, each class was assigned to a different ISI condition. This resulted in 27 students in the 0-day ISI group, 22 students in the 1-day ISI group, and 16 students in the 10-day ISI group.

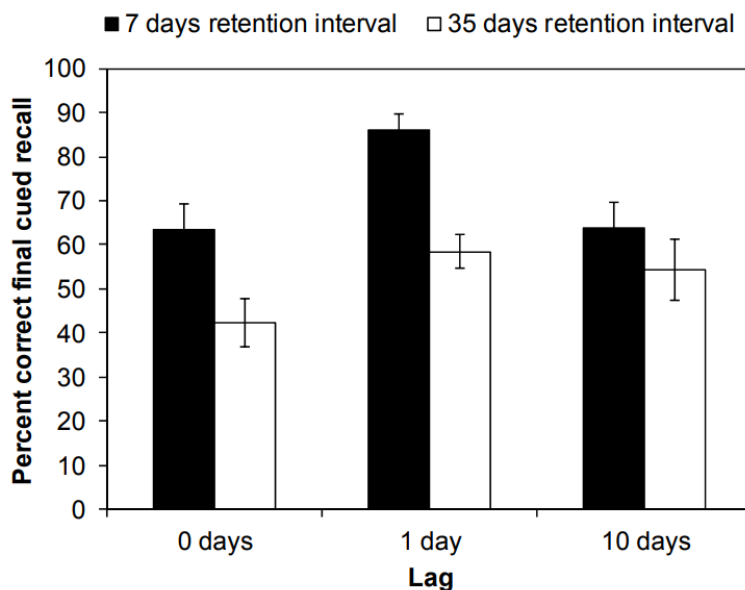
The project consisted of two learning sessions plus a final test occurring after a 7-day RI or a 35-day RI. Later, in each class, students were randomly assigned to the different RI conditions, resulting in overall count of 35 students in the 7-day condition and 30 in the 35-day condition.

The initial presentation of the target vocabulary consisted of a visual projection on the classroom screen showing a word in German and the researcher read it to the class. Two seconds later, the English translation appeared, and the researcher read it out loud. Each pair was shown for a total of eight seconds. Participants were not allowed to take notes. The first learning lesson consisted of two activities and lasted 45-60 minutes. The second learning session took place following the assigned lag and consisted of one activity of 25-30 minutes. The learning activities consisted of the presentation of the German-English target vocabulary, a recognition test, a cued recall test, and a picture quiz. The testing session occurred either seven or thirty five days after the last learning session (depending on the condition). The tests consisted of a multiple-choice test where a German word was provided, and participants had to select the appropriate English translation among three distractors. The productive test required participants to write the English translation given a German word.

The article presents only raw results of the receptive test (multiple-choice test), but the productive-test results that were provided had undergone several statistical procedures. In this review, only the data arising from the receptive test was considered for its simplicity, clarity, and also because this offered a more

straightforward comparison when contrasting this data against other articles in this literature review that also used raw data to analyze results.

Figure 2-3 below shows receptive tests results and indicates that in all of the three lag conditions (0-day, 7-day, 10-day ISIs) participants always scored higher in the 7-day RI test. Most importantly, the figure also shows that for the 7-day RI the highest mean percentages were obtained with the 1-day ISI. This means that at least for this study, for a 7-day RI the ideal interstudy interval was 1 day, which means that the optimal ISI lag was around 14% of the RI. The graphic also shows the inverted-U trend (Cepeda et al., 2008) that matched predictions prior to the intervention, meaning that for a 7-day RI the optimal ISI is one day, otherwise (before or after one day) retention levels might be compromised.



**Figure 2-3: Mean of correctly recalled vocabulary on the final cued recall test as a function of lag and retention interval (Küpper-Tetzel et al., 2014, p. 15)**

Küpper-Tetzel et al. (2014) finally found that, just like in the lab and with university students as participants, vocabulary learning in secondary schools can also benefit

from spaced repetition at the optimal interstudy interval (ISI). Just as predicted, this optimal (ISI) lag is not fixed, and increases as a function of the retention interval (RI).

#### **2.1.4.2 Commentary**

Küpper-Tetzel et al. (2014) is a very interesting paper because it confirmed the spaced repetition outcomes, in line with Bahrick (1979) and Cepeda et al. (2008). Even more interesting is the fact that Küpper-Tetzel et al. (2014) obtained these results in a field study with young L2 learners as participants. In very close agreement with Cepeda et al. (2008), but not with Bahrick (1979), Küpper-Tetzel et al. (2014) found that for a 7-day RI the optimal ISI value is one day (14%) in a receptive test. For the same 7-day RI in a receptive test, Cepeda et al. (2008) suggested a 1.6-day ISI (24%). What is more, based on a similar retention trend to the 7-day RI, in Küpper-Tetzel et al. (2014), for a larger RI (35 days) the appropriate ISI could have been somewhere between one and ten days (2.8% to 28%). For a 35-day RI Cepeda et al. (2008) suggested a 7-day ISI (19%). This shows that both studies agreed on an optimal interstudy interval somewhere around 20% of 35-day RI. This apparent lack of agreement on an absolute value when trying to estimate the optimal RI/ISI combination, still accords with findings from Cepeda et al. (2006), where after reviewing 184 articles on spaced repetition, the authors also saw a wide variation of results. For instance, for a retention interval of 30 to 2900 days they found that the ideal interstudy interval would be somewhere between 29 to 168 days in length.

The paragraph above clearly shows that although there is some agreement between scholars, and RI/ISI findings could serve as reference for future spaced repetition research, there still does not seem to be complete agreement regarding the optimal interstudy interval for a certain retention interval. It is important to remember that the mismatch could also occur since while Küpper-Tetzel et al. (2014) presented a field study with teenagers, Cepeda et al. (2008) introduced an internet-based lab study with mostly adult participants. Therefore, exact findings will not necessarily be similar.

Finally, in Küpper-Tetzel et al. (2014) the 1-day ISI and 7-day RI combination seems to be very evident from the results analyzed (see Figure 2-3), however, it is necessary to note that in the 35-day RI the 1-day lag scored the highest of the three tested ISI

lags. This is important because there is just a minor difference in performance between the 1-day and the 10-day ISI lag, and it is still not clear whether participant performance would have been even higher if there had been a 3- or 4-day ISI conditions tested at 7- and 35-day RI. Another ISI lag condition between the 1-day lag and the 10-day lag would have been a revealing addition to obtain more precise information regarding the ideal ISI lag for a 35-day RI for this study. Another controversial finding arising from this study is the fact that by looking at the 10-day ISI, the highest results were obtained at 7-day RI, instead of at 35-day RI. This is also confusing considering that if the interstudy interval should always be smaller than the retention interval (Bahrick, 1979 and Cepeda et al., 2008) highest scores should have been obtained at 35-day RI.

To sum up, although there are some commonalities, the three papers reviewed so far do not seem to entirely agree on the exact response to the question of when exactly to repeat. Findings from Küpper-Tetzel et al. (2014) partially agree with Bahrick (1979) and Cepeda et al. (2008), but they also present some confounding results highlighting the imperative need for further research regarding the optimal RI/ISI combination. The final paper in this section brings even more interesting results as it differs greatly from Küpper-Tetzel et al. (2014). Lotfolahi & Salehi (2017) presented a spaced repetition field study with primary school children as subjects and is reviewed next.

### **2.1.5 Lotfolahi & Salehi (2017)**

Lotfolahi & Salehi also saw the need to conduct experiments on spaced repetition considering the fact that there is still no final consensus on the ideal combination between the retention interval and interstudy interval (RI/ISI). As opposed to Cepeda et al. (2008) and Bahrick (1979), but in line with Küpper-Tetzel et al. (2014), Lotfolahi & Salehi (2017) investigated the efficacy of spaced repetition to enhance long-term vocabulary retention in an L2 field study with younger learners.

The short study investigated retention of L2 English vocabulary in primary school children in a language school in Iran. The researchers decided to keep the environment as ecological as possible testing whether previous spaced repetition findings would still hold in such a context.

### 2.1.5.1 Summary

The researchers selected 20 English words taken from one of the regular textbooks used in class to be used as target words for the project. Subjects were pre-tested on their knowledge of the words selected to ensure the target words were new to them.

The participants in the study were 28 young learners (from 2<sup>nd</sup> to 5<sup>th</sup> grade) taking beginner L2 English classes at a language institute in Iran. All participants were in two different classes (14 students each) at the language institute.

The research project consisted of two learning sessions, plus a post-test a week after the second learning session, and a delayed post-test five weeks after the second learning session (see Figure 2-4). The intervention extended for six weeks overall and took place during regular class time.

Target words were randomly separated into two lists of ten words each, ten per condition. Each group of participants received a list (20 words) and started learning the words at the same time, but in a different condition. One class began studying list one (ten words) in spaced condition (list 1 = SR) and list two (ten words, split also in two sets of five words each) in massed condition (list 2 = MR), while the other group studied the lists in the opposite order, i.e., list 1 = MR, and list 2 = SR.

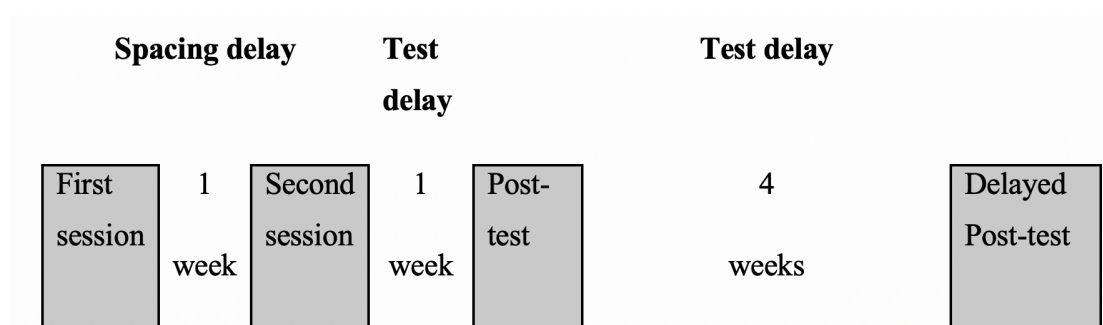
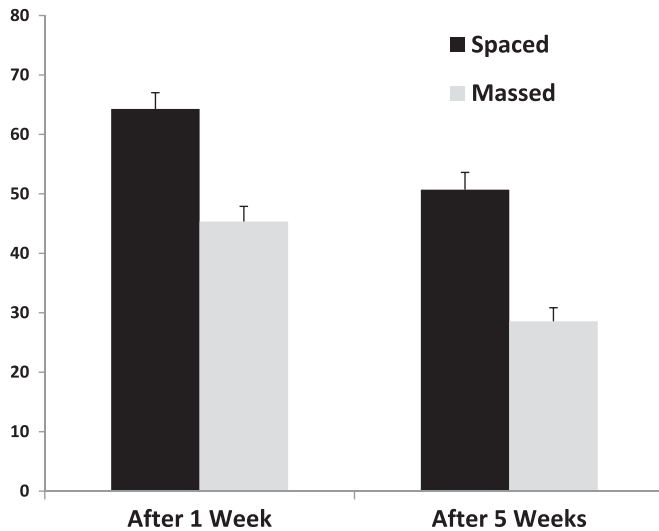


Figure 2-4: Project design (Lotfolahi & Salehi, 2017, p. 4)

In the first session participants learned ten words in the spaced repetition condition and five words in the massed repetition condition. The first study session was firstly divided per condition and it started with spaced repetition and finished with massed repetition for both study groups. Each condition followed four different steps of about five minutes each. In step one, students received a practice booklet, and then all target items were presented to participants, one at a time by their teacher on their class screen. The English word appeared first, then the Farsi translation, and then a sample sentence. Teacher and students read everything aloud. The first step ended with students revising all word pairs on their own. In step two students were asked to open their booklets to page one and work in pairs testing the meaning of the English target words providing the Farsi translation. Step three consisted on working on page two of the booklet having the teacher reading sentences and students repeating after her. This step finished with a 2-minute free rehearsal. The final step was similar to step two, but in reverse order (given a Farsi word, participants provided an English translation). The second study session was similar to the first one, but the conditions were reversed. Participants started working on the remaining five words of the massed repetition condition, and later they worked on the same ten words (studied in session one) of the spaced repetition condition again.

The intervention finished with a test one week after the last learning session, and the same test taken again four weeks after the first final test. The final test consisted of writing down the English translation of the Farsi words given.

Test results of the interstudy interval of seven days (7-day ISI) showed that the mean percentage of items answered correctly in the post-test taken seven days after the last rehearsal (7-day RI) was 64.28% for the spaced repetition (SR) condition and 45.35% for the massed repetition (MR) condition. The five-week delayed post-test (35-day RI) showed different results but with a similar trend: SR: 50.71% and MR: 28.57%. Figure 2-5 below shows percentage comparisons for both conditions in both tests.



**Figure 2-5: Mean percentage of correct recall of massed and spaced items (Lotfolahi & Salehi, 2017, p. 11)**

The authors concluded that although forgetting was taking place, it was lower in the spaced repetition condition for both post-tests. Lotfolahi & Salehi (2017) highlighted the fact that this study tried to keep all research aspects as ecological as possible, using also activities and educational materials typically used by participants.

### **2.1.5.2 Commentary**

At first sight Lotfolahi & Salehi (2017) appears to be a rather short and basic study contrasting spaced repetition to massed repetition. Although this research project does not differ much from previous studies (e.g., Sobel et al., 2011 and Goossens et al., 2012), its findings however deserve some special attention.

Given a fixed 7-day interstudy interval and two different retention intervals (7-day and 35-day) Lotfolahi & Salehi (2017) obtained higher scores in the 7-day than in the 35-day RI test. These findings are at odds with Cepeda, et al. (2006), Cepeda, et al. (2008) and Küpper-Tetzl, et al (2014) that stated that the ideal interstudy interval (ISI) is a portion of the corresponding retention interval (RI), and that the interstudy interval increases at the same time that the retention interval increases. According to those conclusions then, the retention interval and its ideal interstudy interval cannot

have equal values (i.e. in the case of this article: 7-day ISI = 7-day RI). What makes these findings even more interesting is the fact that they agree with findings from Bahrick (1979) in that the ideal interstudy interval is equal in length to the retention interval.

As a consequence, Lotfolahi & Salehi (2017) brought more uncertainty regarding the ideal interstudy interval for a given retention interval. Although many variables could contribute to those findings (e.g., learning tasks, student age, student motivation) this shows that scholars still have not been able to agree on the ideal RI/ISI combination.

To conclude, the results of Lotfolahi & Salehi (2017) suggest that there is a still a need to find ideal combinations between the retention interval and the interstudy interval, especially in long studies. Therefore, a longitudinal study such as the one proposed in this thesis could contribute to the spaced repetition field by testing retention at different retention interval lengths.

### **2.1.6 Conclusion**

After analyzing the articles above, it can only be concluded that more research is still needed to find out when is a good time to repeat according to a certain retention interval (RI). At the same time, findings obtained in spaced repetition lab studies do not seem to translate so easily to field studies. A clear example of this is the lack of practical applications that some findings seem to have when trying to implement them in a real educational environment. For instance, as discussed above, Cepeda, et al. (2008) states that in order to retain information for 35 days, it is best to repeat every seven or eight days. This is in line with a long list of researchers that state that interstudy intervals should be a small portion of retention interval (e.g., Rohrer & Pashler, 2007; Küpper-Tetzel, et al., 2014; Suzuki & DeKeyser, 2017; Serrano & Huang, 2018). If this is to be implemented in an actual classroom in a high-school environment for a whole year, it means that to retain information 35 days after the last learning session, the same information should be repeated about once a week every week for the whole year. At first sight this looks rather monotonous for students and time inefficient considering that students will be repeating the same information once a week rather than learning something new. At least from a practical standpoint,



Bahrick's (1979) findings seem to be more appropriate for ecological environments, considering the fact that Bahrick (1979) suggests repeating every 30 days if information will be retrieved 30 days after the last learning session.

To conclude, the articles just reviewed showed that the optimal combination between the retention interval and the interstudy interval (RI/ISI) is still a work in progress, and more research is needed in order to understand when is best to repeat. However, the RI/ISI combination is not the only factor determining whether information will be retained for a long term or not. The other important factor in retention is learning. If information is not learned properly, it will not be retained properly either (e.g., Bahrick, 1979; Nation, 1990). In line with this, the second section will review different articles that refer to teaching and learning methods to improve retention of information.

## **2.2 Teaching and learning**

Although the previous section could not show exactly when to repeat in order to guarantee the expected retention of information, all of the studies reviewed concurred that the retention interval should be defined first in order to decide how often to repeat (ISI lag). This section, on the other hand, will focus on teaching methods to enhance learning that could eventually lead to proper retention.

This section refers specifically to the different strategies employed by researchers to improve retention of vocabulary and although these articles were selected and grouped together because of their teaching and learning methodologies, they all still refer to spaced repetition and/or retention of vocabulary. This section is further divided into four subsections. The first of the subsections will review a study by Fitzpatrick, Al-Qarni, and Meara who investigated the amount of vocabulary that can be acquired in a certain period of time and how much of that can be retained at different retention intervals. The second subsection discusses field studies that investigate learning and retention with the idea of keeping the classroom as real as possible. The third subsection will introduce new teaching methods that could improve vocabulary acquisition. The final subsection will review Johnson & Heffernan (2006) which deserves a special attention as this study covers the

importance of retaining vocabulary to comprehend authentic materials. This article is replicated in the following chapter, therefore, the justification for replicating it is also explained here.

## **2.2.1 Word gain vs. retention gain**

### **2.2.1.1 Introduction**

This subsection demonstrates that learning and retention are two separate matters, and that retention depends on proper learning, but the fact that information has been properly acquired will not always guarantee long-term retention. The article below (Fitzpatrick et al., 2008) shows that large amounts of cumulative learning at spaced intervals (as opposed to massed repetition) is possible but, at least according to the settings of this study, information is still forgotten quickly.

### **2.2.1.2 Fitzpatrick et al. (2008)**

Fitzpatrick et al. (2008) examined vocabulary learning using spaced repetition (but as opposed to the previous section) putting special emphasis on the number of words that could be acquired in a certain period of time, rather than on the ideal interstudy interval or retention interval. Just like Bahrick (1979), Fitzpatrick et al. (2008) presented a lab study emphasizing the importance of proper learning that could eventually lead to long-term retention.

The researchers decided to conduct a single-subject case study to investigate whether a long list of target words (300) could be learned in 20 days, at a rate of 15 new words a day. The authors decided to use only one subject on the grounds that this project would have been very difficult to implement with more participants. At the same time, the authors also wanted to investigate whether there would be a marked difference between the subject's receptive and productive knowledge of the target words.

#### **2.2.1.2.1 Summary**

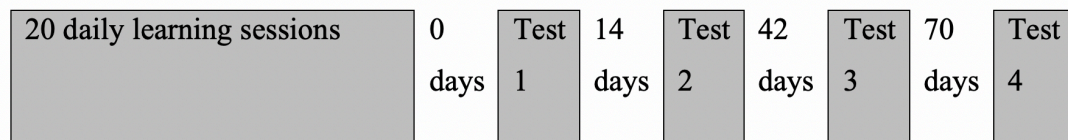
For this study the researchers selected 300 relatively high-frequency Arabic words. Each word appeared on a card in Arabic (spelled using the English transcription of the

Arabic form of the word) together with its English translation. The cards also had 20 numbered boxes to be checked every day the subject studied or reviewed the word. This was done in order to keep count of how many times each word had been rehearsed.

The only subject of the study was an L1 English speaker (Sue), a female 41-year-old teaching linguistics in a UK university. Sue had no prior knowledge of Arabic, except for some basic greetings.

The researchers asked Sue to spend no more than 30 minutes daily learning the words. She was expected to learn 15 new words every day and revise any previously learned words. The authors planned this activity hoping to simulate the vocabulary gain that could be expected to take place in an intensive language course.

Sue was tested four times after the learning sessions were over. Each testing session was divided into two sections eliciting productive or receptive knowledge of the 300 target words at a time. The productive knowledge test was always administered first, and it consisted of a translation task where the participant had to provide the Arabic translation (in Roman script) of an English cue. The receptive knowledge test consisted of the opposite task, i.e. the participant had to provide the English translation of an Arabic word. The first post-test took place immediately after the last learning session, and then two weeks, six weeks, and ten weeks after the first post-test respectively (see Figure 2-6 below).



**Figure 2-6: Study organization, RI days counted from the last learning session (Fitzpatrick et al., 2008, p. 241)**

Basic analysis of test results show that Sue learned almost all of the target words during the learning sessions. The authors had initially thought that the 300 target words would be too high a number of words for Sue to acquire, but that was not the case. Retention results, on the other hand, show that she scored higher immediately after the last learning session, than later in time in both conditions (i.e. receptive and productive). Table 2-5 below shows test scores and percentages between parenthesis. The table clearly shows that recognition scores were higher in every test in comparison to productive recall. For instance, in Test 1 (T1), Sue obtained 286 correct responses (out of 300) in the recognition test, and 283 in the production one. The number of correct responses decreased dramatically towards the last test (T4) in both conditions, especially in productive knowledge.

	T1 (%)	T2 (%)	T3 (%)	T4 (%)
Recognition test	286 (95.33)	262 (87.33)	221 (73.66)	219 (73)
Recall test	283 (94.33)	191 (63.66)	135 (45)	149 (49.66)

**Table 2-5: Number of correct responses per test and percentages (Fitzpatrick et al., 2008, p. 241)**

By analyzing the results more in depth, the researchers found out that, in both conditions, the words learned in the last learning sessions were scoring lower than the words studied in the first learning sessions (the recognition condition had a more gradual decline in comparison to the productive one). In order to obtain more information about motivation and time-on-task the researchers examined Sue's diary and it indicated that starting from session two, for the first 15 minutes she had concentrated on the new words of the day, and afterwards, she had revised previously learned words. The diary showed that Sue had spent 25-30 minutes working on the assigned task each day. Although apparently Sue's motivation had remained constant all through the study, only three of the 15 words acquired in session 20 were answered correctly in the last receptive test (T4). The authors stated that this could

have happened either because of learning overload, or simply because the words learned in the last learning sessions had been rehearsed less.

Finally, the researchers concluded that learning 15 words in 20-30 minutes was not significantly challenging, at least for Sue. The authors also saw that acquiring new words did not become easier as the subject learned more words, and that the subject started to forget words the moment she stopped rehearsing them. A final conclusion the authors made is that receptive knowledge is kept in memory longer than productive knowledge, words that were rehearsed the most were remembered the most, and that shorter words were easier to learn.

#### **2.2.1.2.2 *Commentary***

This study is a clear example that although there could be very good learning, information can still be forgotten quickly. By analyzing retention test results, this investigation supports the claim of the articles reviewed in the first section above (Bahrick, 1979; Cepeda et al., 2008; Küpper-Tetzel et al., 2014; and Lotfolahi & Salehi, 2017) highlighting the importance of the combination between the retention interval and the interstudy interval to ensure information is retained for longer.

Fitzpatrick et al. (2008) demonstrated that, under the conditions set for this study, it is possible to acquire at least 15 new words per session, every day, for a period of 20 days. Using bilingual cards and teaching vocabulary explicitly seemed to be an effective way of acquiring the words easily and quickly, which is in line with the literature (e.g., Thorndike, 1908, Nation, 2007, Schmitt, 2008, Nakata, 2017). Finally, the combination of learning new information and explicit rehearsals of already acquired words in the same session seemed to guarantee proper learning, supporting the claim by Bahrick (1979) that repetition should take place right from the beginning of the acquisition phase.

An important fact to consider is that although results from this study made a significant contribution to the field, its findings cannot be so easily extrapolated to real-classroom settings. For instance, the subject was an experienced highly educated adult who is inherently highly motivated to learn (about) languages. Therefore, the

interest, willingness, and commitment to learning the new words cannot be expected so straightforwardly from younger learners whose motivation and commitment in class will most probably be very different. It does not seem easy to apply such a methodology in a real educational environment, either, where activities should vary to avoid boredom and where student motivation and engagement are key to achieve high learning outcomes.

Finally, the addition of two other tests one day and seven days after the last learning session could have provided more data to contrast these results against those of the studies in the previous sections. This could have provided more data regarding what retention interval obtains the highest retention scores given a 1-day ISI, like in this study. There are three last points that can be extracted from this study. The first one is the fact that retention scores show a decreasing trend from the first test (T1) towards the fourth test (T4). Although there is not enough data, two extra post-tests at shorter retention intervals (between T1 and T2) could have provided more precise information for this particular study regarding retention values at shorter retention intervals (RIs). The second point is that according to deep analysis of individual word scores in the tests, the authors found that Sue could not remember so well the words learned towards the last learning sessions. This appears to be a valid argument for future spaced repetition research since a better option could be to learn all of the target words in session one, and then continue to rehearse all of them in subsequent sessions. This would provide a more balanced exposition to target words. The final point to highlight is that Bahrick's (1979) claim that shorter interstudy intervals show faster and higher cumulative learning, but information is lost faster as well, also seems to apply to this study. This can be seen in the fast decline in test scores as the retention intervals increase.

### **2.2.1.3 Conclusion**

This article is of particular importance for the main project of this thesis for a few reasons. Firstly, it presents an efficient way of learning a large number of vocabulary items in a short period of time, while focusing on retention at the same time. Secondly, a word gain of 15 words per session could be considered as a benchmark in spaced repetition research, but lower learning gains might be initially expected in

field studies with young learners. Thirdly, the strategy of combining learning and rehearsal in the same session probably ensured proper learning and enhanced retention. A concept also shared by Nakata (2017) stressing the importance of multiple learning sessions (specially through the use of retrievals, i.e., recovering information from memory) in long-term vocabulary retention. The final point to consider in Fitzpatrick et al. (2008) is the fact that good learning cannot be correlated with long-term retention. Long-term retention is ensured by both, good learning and the implementation of the ideal interstudy intervals based on the length of the retention interval.

To conclude, considering Fitzpatrick et al. (2008) was a case study, it is imperative to see how the learning process takes place in real educational settings, especially among young learners. Therefore, the following subsection will review field studies focusing on vocabulary acquisition with high-school and primary-school students.

## **2.2.2 Real ecological teaching and learning**

### **2.2.2.1 Introduction**

The articles in this subsection seek to bridge the gap between vocabulary acquisition in lab condition and field studies. Considering that there are many variables that could affect results (e.g., student engagement and motivation) this subsection will focus on research studies that aimed at leaving the teaching and learning environment as real as possible. This is done in order to simulate actual classroom intricacies and obtain a clearer idea of how to best implement spaced repetition (SR) in real educational environments.

There were several attempts to introduce spaced repetition in everyday classrooms varying in methodology, duration, and actual spaced repetition integration into the curriculum. Some differences that can be seen in the articles reviewed below are found specially in whether the target items were part of the syllabus and were to be tested later as a course requirement (as seen in Bloom & Shuell, 1981), to techniques and activities used to teach students. Bloom & Shuell (1981), for instance, employed longer interstudy intervals as they considered that those intervals resembled actual classroom situations. Sobel et al., (2011) also aimed at ecological environments, but

the researchers of the study decided to use actual teachers to teach the students. Finally, Goossens et al., (2012) applied a variation of the activities that participants were required to work on considering that in everyday classrooms students do not work on the same exercises all the time.

#### **2.2.2.2 Bloom & Shuell (1981)**

This study by Bloom & Shuell represents historically one of the first attempts to bring spaced repetition from the lab to actual school settings and using high-school students rather than university students as subjects. The authors concentrated on traditional retention research contrasting spaced repetition (SR) to massed repetition (MR) and used students learning French as a foreign language. The researchers did not modify the curriculum and investigated the importance of the length of the interstudy intervals (ISI).

To start with, the researchers analyzed previous studies (e.g., Underwood, 1961; Houston & Reynolds, 1965; Houston, 1966) and they saw that those scholars were using shorter interstudy intervals of one to four minutes for spaced repetition, and two to eight seconds for massed repetition, but it was through longer lags (interstudy intervals) that the benefits of spaced repetition seemed to appear more prominently. For instance, the authors found that in Keppel (1964) participants learning through spaced repetition retained similar amounts of material after 29 days (34%) in comparison to the massed repetition group that retained 31% after 24 hours. Therefore, this showed that the length of the intervals between learning sessions was a key factor for higher retention levels using spaced repetition. In line with this trend, Bloom & Shuell (like Bahrick, 1979) also opted for longer intervals in their study.

##### **2.2.2.2.1 Summary**

The target words used in Bloom & Shuell (1981) were 20 French vocabulary items representing names of occupations and their English counterparts (e.g., *l'avocat* - 'lawyer'). These words were part of the required vocabulary for the regular class course and would also be part of the test at the end of the unit. All 20 word-pairs were printed on a sheet of paper and given to the students to work with (only during class) in preparation for a vocabulary test to be given at the end of the week.



The participants were 52 high-school students attending three different classes at the school where the project took place. Subjects were in grades 9-12 in the American system (ages 15 to 18) taking French as a foreign language, and they were randomly split into two groups (the massed repetition group and the spaced repetition group).

The experiment was conducted as part of the regular, ongoing activities of each class. It consisted of a series of three learning sessions, an announced test following session three (0-day RI), and an unannounced test four days after that (4-day RI). To keep subject motivation, participants were told that the target words were part of the curriculum and would be tested at the end of the unit.

Both groups had a total of 30 minutes to work in class with the 20 words. These 30 minutes were separated into three different 10-minute activities. These activities consisted of a series of three, 10-minute written exercises purposely created for the study. The first activity was a written, multiple-choice exercise (e.g., *fireman: le proviseur, le facteur, le pompier*) where the students had to select the correct option, and, if they were wrong, they had to correct their own mistakes using the word list mentioned before. For the second activity students were given sentences in French with the description of an occupation and they had to provide the occupation in French (e.g., // *cultive les legumes et les fruits*). The last activity consisted of a written test that served as practice for the final test at the end of the intervention. In the test learners were asked to write the French word for each occupation provided in English (e.g., *businessman* ). The group under the spaced repetition condition worked on the project for ten minutes each day for three consecutive days. The massed repetition group worked for 30 minutes on the same day the spaced repetition group had the last learning session.

Both groups were tested immediately following the third study session, and they took the same test again four days later. Both tests consisted of a written activity where participants had to provide the French translation of the occupations given in English.

Results collected from the first test showed that there was similar vocabulary learning uptake for both conditions. Data from the second test (4-day RI), however, revealed that after four days from the last study session forgetting was already taking place.

Although both conditions scored lower in the second test, participants in the spaced repetition condition (SR) managed to retain more words than those in the massed repetition (MR) group. This agrees with findings from Bahrick (1979). Table 2-6 below shows mean percentages for both tests with standard deviation between parentheses.

Group	Test	
	Test 1	Test 2
SR	84.25 (3.00)	75.2 (3.78)
MR	80.60 (2.64)	55.75 (4.02)

**Table 2-6: Tests mean percentages with standard deviation (Bloom & Shuell, 1981, p. 246)**

The researchers found that, as seen by tests results, spaced repetition in a school environment can significantly increase students' retention of previously learned knowledge. The authors also stressed the fact that spaced repetition can be used in conjunction with everyday classroom methodologies to improve student memory, and not only in language courses. Bloom & Shuell (1981) concluded that learning can be achieved quickly with short interstudy intervals, and that information can be later maintained employing longer interstudy intervals. The authors finally added that the benefits of spaced repetition are not seeing during the learning period, but they become evident during long-term retention.

#### **2.2.2.2.2 *Commentary***

Findings from this study show that the primary goals of the researchers were met. Spaced repetition could be implemented into a language course with positive results, and retention tests are consistent with previous studies contrasting spaced repetition and massed repetition (e.g., Keppel, 1964) showing that spaced repetition scores are higher than those of the massed repetition condition in long-term vocabulary retention

tests. Also, the authors saw that L2 vocabulary can be taught explicitly and successfully to enhance retention. Finally, and in line with Bahrick (1979), Bloom & Shuell (1981) also found that longer interstudy intervals (as opposed to short intervals) are more appropriate for long-term retention.

At the beginning of their study Bloom & Shuell mention that previous researchers had used very short interstudy intervals (ISI). Therefore, the authors decided to implement a larger interstudy interval of one day in their study. Although the 1-day ISI used in Bloom & Shuell (1981) is longer than interstudy intervals of previous research, it still appears to be rather short. The study could have benefited from adding other longer ISI conditions and test retention levels afterwards to further explore the effect of the length of the interstudy interval in long-term vocabulary retention. In this line and seeing that longer lags seem to resemble what happens in real classrooms, Sobel et al., (2011) investigated the effects of a seven-day interstudy interval over long-term vocabulary retention. The researchers of the study decided to leave all conditions as ecological as possible to test retention while also using learning activities students typically work on in the classroom. The article is reviewed below.

### **2.2.2.3 Sobel et al. (2011)**

Like Bloom & Shuell (1981), Sobel et al. (2011) also investigated the efficiency of spaced repetition to enhance long-term retention of information while simulating closely everyday classroom situations. The study contrasted the efficacy of spaced repetition (SR) against massed repetition (MR), and takes place in an actual L1 classroom, with actual teachers, having middle school students as subjects.

Sobel et al. (2011) stated that teachers in general are always trying to find ways to improve what students can remember and that spaced repetition could help in that sense. Despite the fact that some researchers had conducted field studies employing spaced repetition (e.g., Bloom & Shuell, 1981) Sobel et al. (2011) still argued that there was a lack of implementation of spaced repetition in real classrooms. In line with other scholars (e.g., Dempster, 1988), the authors of the study explained that spaced repetition still needs to be applied to real-world classrooms and more field studies are needed. In response to that, the researchers conducted a field study

focusing specially on middle school children since they saw lack of spaced repetition research with them. Also, like Bloom & Shuell (1981), the authors found that most spaced repetition studies with children used very short interstudy intervals. Therefore, through this study the researchers aimed at introducing retention intervals of a certain length (that they considered were more relevant to real educational environments) and larger interstudy intervals. The researchers considered that longer gaps between learning sessions resembled actual classroom learning situations and could contribute better to long-term vocabulary retention.

To leave classroom conditions as ecological as possible, the researchers had the actual class teachers leading the sessions using teaching and learning methods students were used to. Considering younger learners might take extra time to get ready to work, the authors provided additional time for preparation, handing out and collection of materials during every session.

#### ***2.2.2.3.1 Summary***

The overall target words used in the study were eight (four per condition) English words (judged by the experimenters to be new to fifth-grade children) that were arbitrarily selected from a GRE word list. The GRE (Graduate Record Examinations) is a standardized test that is required for admission by most graduate schools in the United States. The researchers made sure that all target words fell outside the most frequently 9000 words (e.g., tacit, edict, gregarious, coerce) in spoken and written English according to WordCount.org (<http://number27.org/assets/misc/words.txt>). They considered this range of 9000 words was outside the words that typical fifth-graders would know.

The subjects were 39 middle school participants from two different fifth-grade classrooms averaging ten years of age. The project employed an intact-cohort method, therefore, each class represented a different group in the study. Also, looking for an ecological environment, the researchers decided that participants would be taught by their usual classroom teacher.

All participants went through two experimental conditions massed repetition (MR) and spaced repetition (SR), learning four words per condition. The massed repetition condition consisted of two 15-minute learning sessions separated by less than a minute. In the spaced repetition condition there were two 15-minute learning sessions separated by seven days. In each learning session, five minutes were used for instructions, handing out and collection of materials. There were ten minutes of proper learning.

At the beginning of every session participants received a three-page booklet according to the condition they were working on (MR or SR). The first page had four target words to be learned, the second page had the definitions of those words, and the third page provided a space for participants to write the definition of the words and use them in a sentence.

The session started with the teacher presenting the target words to participants with an overhead projector. Subjects were instructed to read along with their teacher who went over the assigned target words, their definitions and sample sentences. Then subjects were asked to turn to page one of the booklet and write the definitions of the four target words on the page. Later participants turned to page two and were required to review the definitions by themselves. Finally, participants turned to the last page of the booklet, wrote down the definition of each one of the target words, together with a novel sentence containing each word.

The second learning session took place as planned, either one minute after the first one in the massed repetition condition, or one week after the first session in the spaced repetition condition. The second session was an exact replication of the first one. In each learning session the teacher always monitored the students to ensure they were on task.

Participants went through two final exams (one for each condition) five weeks after the second learning session. They had ten minutes per exam to write the definition of the four target words provided according to each one of the different conditions.

Data was collected from session one in each condition and also from the post-test taken by both groups five weeks after the last learning session. In order to check participants' performance during session one the authors collected data from student responses to find that both groups were almost at the same exact level. After the post-test, the researchers collected results and conducted a paired sample *t*-test showing mean percentages of spaced repetition condition:  $M = 20.8$ , and  $M = 7.5$  for the massed repetition condition ( $t = 3.0$ ;  $d = 0.48$ ;  $p = .004$ ). This means that participants remembered three times as many definitions in the spaced repetition condition as they did in the massed repetition condition, with a highly statistically significant difference between the groups (as reflected in a  $p$  value  $<0.05$ ), and a small size effect (as reflected by a  $d$  value  $<0.5$ ).

After analyzing data, the authors concluded that, as shown in a 5-week RI post-test, the 7-day gap (spaced repetition condition) showed greater gains in comparison to the massed repetition condition. Sobel et al. (2011) reported that those findings could easily generalize to actual classroom situations in middle schools. The researchers stated that spaced repetition could be easily integrated into the curriculum using educationally relevant interstudy intervals and retention intervals. The researchers finally claimed that with a quick reorganization of the lessons, and without increasing teaching time, spaced repetition could be beneficial to improve results in course tests.

#### **2.2.2.3.2 Commentary**

Again, like Bloom & Shuell (1981), Sobel et al. (2011) demonstrated that not much is needed to integrate spaced repetition into the curriculum, and although this is an L1 study, its findings can easily be transferred to L2 vocabulary acquisition. For an easy integration the researchers used the actual teachers to lead the sessions, they employed teaching and learning methods students were used to, and even allocated extra time for preparation, handing out and collection of materials during every session (which demonstrated good knowledge of real classroom intricacies).

There are some important issues that still need to be considered. To start with, the total number of target words (eight), was very small, and the study appears to be rather short with only two learning sessions. More revealing data could have been

obtained if the study had been longer (more learning sessions) to discover how this age group could have responded, especially in matters of concentration and/or motivation in the long run. Secondly, data collection seems to be rather confusing. Instead of collecting and analyzing data from session one and then contrasting it against post-test results, it seems more straight forward to run a pre-test prior to the study and then contrast results against the post-test. This way it would have been more precise to check participant knowledge before the intervention instead of simply assuming it to be zero. Finally, although the spaced repetition condition scored considerably higher than massed repetition, an average SR group score of 20.8% in the retention test seems to be extremely low, showing that may be something went wrong along the way in the study, but this is not mentioned by the authors.

To conclude, Sobel et al. (2011) succeeded in embedding spaced repetition into a middle-school class while keeping many aspects of teaching and learning as ecological as possible. This study further contributed to the fact that spaced repetition seems to be compatible with an ongoing curriculum and it could be integrated seamlessly.

Up to this point this subsection has demonstrated that good learning should be combined with appropriate implementation of spaced repetition in order to enhance long-term retention, and lab findings still hold for field studies and across different age groups. The last study in this subsection (Goossens et al., 2012) aimed at expanding Sobel et al. (2011) by introducing more target words, more learning sessions, two different retention intervals, and also by claiming that learning activities should be different to improve learning.

#### **2.2.2.4 Goossens et al. (2012)**

In the same line as Sobel et al. (2011), Goossens et al. (2012) also investigated how spaced repetition could be applied to young learners in an ecological L1 study simulating closely everyday classroom situations. As opposed to Sobel et al. (2011), however, Goossens et al. (2012) used even younger learners (primary school students) as participants working on different learning tasks investigating whether the same results would still apply to them.

Goossens et al. (2012) found that only a few spaced repetition studies used primary-school learners, and the majority of them employed practices that were uncommon in the actual classroom. For example, the authors saw that previous research employed very short retention intervals, that, according to them, are not common in real educational settings. The researchers also noticed that Sobel et al. (2011), for instance, used the same activities across different learning sessions. Goossens et al. (2012), however, argued that there should be a variety in the learning tasks in order to improve learning and retention. The researchers of the study believed that through a variety of activities they could increase engagement, avoid boredom in the students, and eventually improve learning. Therefore, the authors introduced different kind of learning exercises in their study to test long-term retention in two different conditions (spaced repetition and massed repetition).

#### **2.2.2.4.1 Summary**

Goossens et al. (2012) used 30 target words taken from current grade four learning materials (not known to subjects since they were in grade three). The words were randomly split into two groups of 15 words each, and later arranged into thematic sets of five words each. In order to study the words, there were three different activities participants were required to work on. The first activity consisted of a fill-in-the-blanks task where subjects had to complete a phrase provided using one of the target words. In the second activity participants had to answer true-or-false to statements about the word and their definitions. The last activity was a multiple-choice task where subjects had to select the appropriate word according to the definition provided.

Participants were 33 grade-three students from two different classes in a primary school in the Netherlands. Average age of subjects was 8.9 years.

Goossens et al. (2012) introduced an intact-cohort research method making up two groups out of two different classes. The 30 target words were divided in half and each set was assigned to either a spaced repetition or a massed repetition condition. As opposed to Sobel et al. (2011), the classes were taught by the experimenters, not the usual classroom teacher.



The study was organized into four classes, one per day on four consecutive days. There was a general introduction session (covering all 30 target words) and three learning sessions. There was a post-test one week after the last learning session, and one delayed post-test four weeks after the first test. Both study groups followed the same format, but the word lists were studied in the opposite condition (while one group was learning 15 words in one condition, the other group was learning the same words in the opposite condition).

During the initial session the researchers introduced all 30 target items to participants. This was followed by three different learning sessions (either through spaced repetition or through massed repetition). The first presentation (of all target words) took place visually through a class projector, orally by teacher and students going over the meaning of the words together. Then, the learning sessions took place according to each condition. In every session, participants went through both conditions, studying all 15 words in the spaced repetition condition list every time, and a different set of five words in the massed repetition condition list per class. In each SR learning session, after the general presentation, participants learned all of the assigned words working on just one activity type. In the following SR session, a different activity type was assigned. For instance, in session one participants studied all of the assigned words in a fill-in-the-blanks task, while in the next session they studied all of the pertaining words in a true-or-false task. In each MR session, on the contrary, participants learned the five words assigned for the day going through all of the three different activity types. Participants received feedback from the researchers in every activity, so they could evaluate their individual performance.

Data was collected through the post-test a week after the last learning session and through the second test four weeks after the first test. Both tests were identical and consisted of a productive activity where subjects were required to write down the correct word of a given definition.

To test that learning was taking place evenly in both conditions, the researchers graded every single exercise in the learning sessions. Overall results showed that there

was no significant difference between conditions (MR:  $M=86.81\%$ ,  $SD=14.68$ ; SR:  $M=87.33\%$ ,  $SD=15.39$ ) and that both groups learned equally well.

Results from the retention tests showed a different outcome (see Table 2-7). Spaced repetition outperformed massed repetition in both post-tests, but the greater difference was found in the 1-week RI test. This means that at 1-day ISI lag, participants managed to retain more information (one week after the last learning session) when target words were learned at spaced intervals rather than when they were massed.

Condition	Final test	
	After 1 week	After 5 weeks
MR	46.46% (25.85)	42.22% (23.07)
SR	55.96% (26.24)	49.49% (27.13)

**Table 2-7: Mean percentage of final tests for both conditions, plus SD (Goossens et al., 2012, p. 969)**

The researchers concluded that, as they had expected, participants scored higher in the retention tests when they learned the target words through spaced repetition. The authors argued that, as shown in the study results, spaced repetition benefits also applied to younger learners in a field study and that using different activity types improved learning. The researchers also recognized that apparently, participants lost some concentration towards the end of the lessons when massed learning took place, and this could have affected overall massed repetition scores in a negative way.

#### **2.2.2.4.2 Commentary**

The final article in this subsection is in line with the two previous studies and it also shows that the researchers managed to obtain positive results in retention scores (in the SR condition), and that they managed to successfully integrate spaced repetition in the classroom. The extra aspect that Goossens et al. (2012) included (in comparison to

the other two articles) is the use of a different learning activity in every learning session, which apparently also contributed to the retention results.

Goossens et al. (2012) compared their findings against those of Sobel et al. (2011) and stated that their project produced more realistic findings (than Sobel et al., 2011). Goossens et al. (2012) highlighted the fact that the study itself resembled real educational settings by using different activity types for learning, and that retention scores were significantly higher than those of Sobel et al., (2011). Although this is not entirely wrong, there is an interesting consideration to be made. Goossens et al. (2012) used four learning sessions while Sobel et al. (2011) only used two. Arguably, this could have given participants in Goossens et al.'s (2012) study more opportunities for learning. This would be in line with Bahrick (1979) and Fitzpatrick et al., (2008) who stated that more rehearsals contribute to better learning. Arguably, this could have also enhanced long-term retention.

This difference in test scores that Goossens et al. (2012) refers to, seems to be of particular importance also, since this study showed that mean scores (49.49%) in the 35-day RI at 1-day ISI were much higher than Sobel et al.'s (2011) mean scores (20.8%) in the 35-day RI at 7-day ISI. All things being equal, according to Cepeda et al. (2008), Sobel et al. (2011) should have obtained higher retention scores. Therefore, this difference in scores among those two studies contradicts Cepeda et al.'s (2008) findings stating that the ideal interstudy interval for a 35-day RI is an interstudy interval of about eight days for productive knowledge. This could suggest (among many different variables that could also affect learning and retention) that primary-school children may retain information in a different way, that adding more learning sessions contributes positively to learning and retention, that learning through different activities can indeed boost learning, or even that after long years of research there is still much to be learned about spaced repetition and retention of knowledge. This, as a consequence, suggests that further research in spaced repetition is needed to address those issues.

All in all, findings from Goossens et al. (2012) contributed to the field of spaced repetition and long-term retention of information in that they also saw the benefits of

spaced repetition in younger children. Although the authors reported certain lack of concentration towards the end of some lessons, the inclusion of different activities across the learning sessions seemed to have contributed to the higher SR scores in retention tests.

### **2.2.2.5 Conclusion**

The three articles reviewed in this section all aimed at bridging the gap between lab and field studies. It is interesting to see that they all used different age groups as participants, and spaced repetition findings still seemed to hold in all situations. All three studies also successfully integrated spaced repetition into a real classroom leaving as many aspects of teaching and learning as ecological as possible, demonstrating that spaced repetition can be implemented seamlessly into everyday classrooms.

As pointed out before, Bloom & Shuell (1981) noticed that in spaced repetition research there was a lack of field studies implementing spaced repetition in the classroom. The same article also pointed out that prior research used interstudy intervals that were too short or that subjects used were mainly adults or young adults. Several years later researchers still continued to highlight the same issues (e.g., Cepeda et al., 2006; Goossens et al., 2012). This clearly shows the need to continue to conduct spaced repetition field studies that could help understand how retention of information actually works, and to find how to best implement spaced repetition in real educational settings.

Although this subsection showed that spaced repetition could be integrated leaving classroom settings almost untouched, there is still another component that needs to be considered. This is the case of innovative teaching methods that look at improving learning and increase motivation and engagement. Therefore, the two final subsections will cover studies that introduced innovative teaching methods aiming at faster and better vocabulary learning while still trying to avoid forgetting.

## **2.2.3 Innovative teaching methods:**

### **2.2.3.1 Introduction**

The previous subsection focused on bridging the gap between the lab and real educational settings, trying to implement spaced repetition seamlessly into the everyday classroom. As shown by the three articles reviewed before, the benefits of spaced repetition found in lab studies, can also be found in field studies without making major adjustments to the curriculum or to everyday teaching methodologies.

This subsection, on the other hand, concentrates on studies that purposely modified everyday teaching methods. Therefore, these research projects introduced research-based methods that could help learn large amounts of vocabulary in a short period of time while also contributing to long-term retention.

The first of the articles reviewed (Erbes et al., 2010) is unique on its own since it also reflects on the difficulties that could arise when introducing memory research into a real classroom. The remaining three articles in this subsection (McLean et al., 2013; Milliner, 2013 and Gryzelius, 2016) will discuss how the introduction of explicit teaching through online flashcard programs can improve vocabulary acquisition.

### **2.2.3.2 Erbes et al. (2010)**

Although Erbes et al. (2010) did not focus strictly on spaced repetition, the article still investigated Spanish vocabulary learning through new teaching methods, aiming at long-term retention. The study by Erbes et al. (2010) examined whether vocabulary acquisition based on memory research can be integrated into a real classroom, and how information is retained over a period of time.

This article focuses on the fact that brain research could be applied to everyday classrooms, but in order for this to be successful, there are several issues that need to be addressed beforehand. The researchers of the study focused specially on the practical aspects that need to be considered for actual successful integration. The researchers investigated how educational psychology along with L2 teaching and human memory research can improve learning Spanish as a foreign language in k-12 (from kindergarten to grade 12 in the American schooling system) schools.

Erbes et al. (2010) was based on the ideas of Caine & Caine (1991) stating that language learning should focus mostly on what the authors refer to as ‘elaborative teaching approaches’. These methods make use of the full potential of learners’ brain, as opposed to simple traditional methods (the authors define *traditional* as the methodology used every day in classroom settings by teachers). An example of a traditional method is rote repetition, and it consists on memorization of information based on repetition. Erbes et al. (2010) reported that Caine & Caine (1991) found that elaborative methods, as opposed to traditional ones, allow learners to extract meaning from new information and to connect it to familiar material. This way, more associations lead to easier retention and retrieval.

#### **2.2.3.2.1 Summary**

The target words for this study were 30 different food nouns in Spanish. Participants received two lists of 15 words each, a worksheet containing 15 pictures of the items. In the innovative teaching method teachers used real food to teach the lesson.

The project involved 78 students (native English speakers) from two different public high-schools in the USA taking Spanish as L2 at the beginner level. There were 42 students in school A, and 36 in school B (male and female) and their age mean was 14 years old.

Test instruments consisted of a pre-test, two scripted lessons, and 6 post-tests. The teacher in school A was an experienced female teacher in her mid-50’s, while the teacher in school B was a male in his 20s with very little teaching experience.

Both teachers followed scripted lessons to ensure they were following the same methodology in their classes. There were two phases in the study, in each phase, the teacher followed either the traditional or the innovative teaching method, and then changed the teaching method in the following phase. In each phase participants had to learn 15 Spanish nouns. For the traditional method the teacher read the words in Spanish twice, and then twice in English in front of the class, and students repeated after the teacher every time. Afterwards, the teacher modeled the pronunciation of every word in Spanish again, and students repeated along. For the last activity,

participants were given a picture of each one of the 15 Spanish nouns and they were required to write the Spanish words below each corresponding picture.

The non-traditional method involved deeper processing as the learners associated the sounds to real objects presented to them, followed by an indication of pleasant or unpleasant connection to the objects. Participants were given a list of 15 items, and then the teacher showed a real food item of each one of the words in the list at a time. Later the teacher said the word out loud in Spanish to the class, and the students repeated the word every time. Afterwards, participants were given a worksheet with pictures of food items. For each item, the teacher pronounced the word in Spanish and the students repeated immediately after. Finally, learners wrote the word in Spanish next to its picture and checked a box next to it saying *me gusta* ('I like it') or *no me gusta* ('I do not like it').

After each lesson participants were quizzed three times. The first test took place immediately after the lesson to test recollection of target vocabulary. The second test took place three days after the lesson testing retention, and the third test took place 24 days after the lesson testing long-term retention. Each test consisted of 15 questions where participants had to write the Spanish words below their corresponding picture.

Tests results showed that for School A (see Figure 2-7) the non-traditional method scored higher than the traditional method in the immediate test, but performance decreased with time. The traditional method, on the other hand, showed more homogeneous results across all three tests. The scores of the first test of the traditional method were lower than the innovative method. The traditional method showed slightly lower results in the third test, although still higher in comparison to the non-traditional method. It is important to highlight that for some reason participants scored higher in the immediate test using the non-traditional method, but the traditional method, instead, seemed to be more effective for long-term retention.

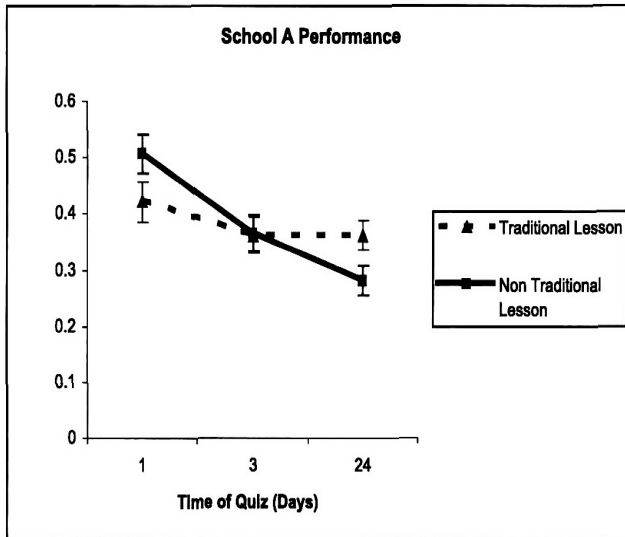


Figure 2-7: Memory effect for school A (Erbes et al., 2010, p. 125)

School B showed different results altogether (see Figure 2-8). The non-traditional method scored higher across all three tests in comparison to the traditional method. In school B there is also a steep drop in scores towards the third test in both methods.

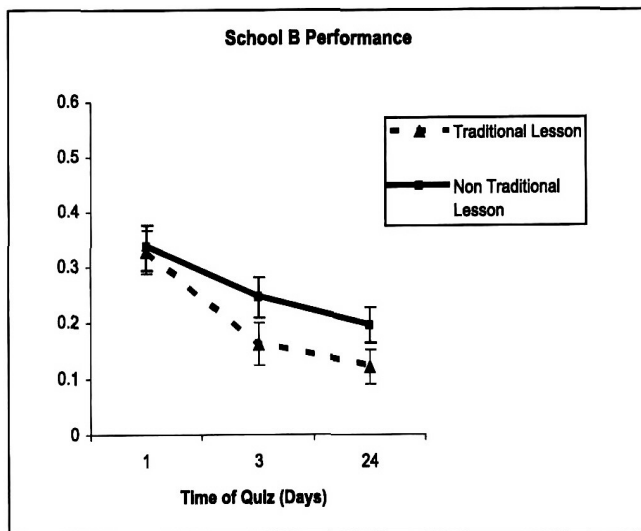


Figure 2-8: Memory effect for school B (Erbes et al., 2010, p. 126)



Both figures above show that after a single exposure memorization of target words decreased over time. School A scored higher in all aspects in comparison to School B. The authors concluded that even though both teachers followed the same scripted lessons, there were several variables that could have led to such results, e.g. classroom management and student-teacher relationship. The authors finally highlighted the fact that motivation, or lack of it actually, could have also affected results. The researchers stated that this had probably not been an issue in many studies on these topics in the past. The reason for this seemed to be the fact that those studies had taken place at the university level where students were rewarded with prizes or credit. As a result, the authors suggested that future research on these topics should use high-school participants to understand how this age group reacts to the implementation of memory research in educational settings.

#### ***2.2.3.2.2 Commentary***

Erbes et al. (2010) referred directly to one of the main issues discussed in the Introduction chapter: how to teach vocabulary to ensure proper retention. The study found that even when a certain method succeeds in a lab study, it does not mean that the same findings can be replicated in a field study straightforwardly. Therefore, certain variables pertaining to classroom intricacies should be considered carefully before conducting such an experiment. For instance, the study highlighted that age group can be a factor for engagement, and teacher-learner relationship might also influence the outcome of a field study.

A major flaw of this article seems to be teacher selection. Clearly participants in school B did not learn at the same level as those in school A (as shown by the results of the first test). This stresses the fact that, at least in the context of this study with only one learning lesson and in order to contrast retention levels in different groups, both learning levels should have been equal. That is to say, if there is poor learning, retention levels are also going to be low. Therefore, the researchers should have ensured that both groups learned at the same level. Arguably, then, one of the reasons why findings were not what the researchers had expected, is that they selected two

teachers with very different backgrounds and styles to lead the lessons. That seemed to have caused strong differences in participant engagement, and consequent learning.

To conclude, this study served to highlight two important issues. To start with, when conducting educational field studies external variables such as teacher selection, learner engagement, and learner motivation should be carefully considered as they can easily influence research results. Another issue that arises from this study, is the fact that it stresses what other scholars had stated before (e.g., Bahrick, 1979; Bloom & Shuell, 1981; Fitzpatrick et al., 2008), learning and retention are two different processes that, although they are closely connected, they are separate. If there is no proper learning, retention levels are going to be low. Even if learning is high but information has not been spaced properly (as in the case of massed repetition) long-term retention can still be poor.

The three articles reviewed next address specially two of the issues mentioned above that could greatly influence learning and field research outcomes: student engagement and motivation. As a result, the researchers listed below decided to teach vocabulary explicitly through technology to promote learning while keeping participants engaged and improve motivation.

### **2.2.3.3 McLean et al. (2013)**

Erbes et al. (2010) revealed a crucial aspect in vocabulary acquisition field research, that is the fact that if learners are not engaged, results will be negatively affected. This can be really problematic in longitudinal studies extending over months considering the fact that learners will need to be highly motivated to actively participate in the tasks over longer periods of time. With this in mind, McLean et al. (2013) introduced a vocabulary acquisition field study for a whole year using a digital flashcard website. The researchers in the study believed that through the use of the platform (as opposed to working on paper), they could increase learners' vocabulary sizes while also keeping them motivated.

The researchers combined the digital platform with regular classwork and homework activities aiming at increasing participants' vocabulary sizes. The project focused

strictly on learning, making use of both implicit and explicit methods, while leaving the spaced repetition aspect to the online system.

McLean et al. (2013) was based on Elgort (2011) claiming that explicit learning is an efficient method to acquire vocabulary quickly. McLean et al. (2013) investigated whether the use of digital flashcards could improve participants' scores in the VST (Vocabulary Size Test) test by Nation & Beglar (2007). The VST is a test that was designed to measure the receptive vocabulary size of learners of L2 English.

#### **2.2.3.3.1 Summary**

The researchers used implicit and explicit learning activities to improve learners' vocabulary knowledge. For implicit learning they assigned graded readers to be read at home by two of the groups of the project (there was a third group in the study, with no assigned reader). For explicit learning the authors introduced an online digital flashcard program. The researchers opted for digital flashcards since they considered that flashcards on paper had many limitations while digital ones could bring more opportunities for learning.

The online flashcard platform used in the study was WordEngine (<https://www.wordengine.jp/>). This is an online flashcard program to learn vocabulary recording learner study-time, number of words studied, and words acquired. Apart from the flashcards, the system also offers games to reinforce vocabulary learning, and in order to help with retention, it uses an automatic spaced repetition method with an expanding interstudy intervals.

The subjects were 182 Japanese university students taking compulsory English courses. At university, students were taking two 90-minute classes of L2 English a week (one receptive and one productive).

This study took place during the productive class alone, once a week for 90 minutes during the whole academic year (28 weeks), plus weekly assigned homework. Participants took a VST test at the beginning of the school year as a pre-test. Then, all participants took also a test using WordEngine, which uses a series of databases and corpora to automatically determine a learner's vocabulary knowledge. Later, the

system determines what words each learner needs to learn based on the responses. The system then presents words to users based on their individual level, therefore, not all participants learned the same words nor in the same sequence. The system has two courses, and after the automatic placement, participants were automatically assigned to either the basic or to the advanced course.

The researchers started by dividing participants into three groups: there was a control group, and two experimental groups: vocabulary + ER (extensive reading), and vocabulary. The activities participants had to go through were also divided into in-class and out-of-class. Out of class, the control group had to read a graded book per week. The vocabulary + ER group, on the other hand, had to read a graded book plus spend an hour on WordEngine a week. The vocabulary group had to complete two hours on WordEngine per week, with no reader assigned. In class, all participants learned vocabulary incidentally with activities focused on productive knowledge. The control group was asked to write a review of the weekly assigned reader and retell the story to other subjects. The vocabulary + ER group had to work on a website and answer questions based on the weekly assigned reader. The vocabulary group had to work on any assigned activity to improve their oral and written production. Table 2-8 below shows a summary of assigned activities per group. In class, teachers wandered around checking that students were working on the appropriate tasks. To ensure subjects from the experimental groups spent the allocated time on the flashcard platform, researchers checked time-on-task per participant using the managing tool of the program.

<b>Group</b>	<b>In-class activity</b>	<b>Out-of-class activity</b>
Control	Retelling and writing based on ER	One Penguin reader weekly
Vocabulary + ER	Oral and written discourse activities	One reader a week plus an hour on WordEngine
Vocabulary	Oral and written discourse activities	Two hours of WordEngine a week

**Table 2-8: In-class and Out-of-class activities per group (Mclean et al, 2013, p. 87)**

Raw data from VST scores shows (see Table 2-9) that previous to the study the vocabulary group had the highest level of English, and it was the one scoring higher also in the post-test. Both experimental groups scored higher in the post-test (as expected by the researchers). What was surprising however, was the fact that the vocabulary + ER group obtained the higher vocabulary gain. The authors first believed that more time spent on WordEngine would cause the highest gains.

<b>Group</b>	<b>N</b>	<b><i>M</i> Pre-test</b>	<b><i>M</i> Post-test</b>	<b>Pre-post Difference</b>
Control	57	2570.18	2645.61	75.44
Vocabulary + ER	61	2360.66	3508.20	1147.54
Vocabulary	64	3214.06	4321.88	1107.81

**Table 2-9: Vocabulary size mean of pre and post-test scores (McLean et al, 2013, p. 91)**

The authors finally concluded that the inclusion of incidental learning through an online flashcard system helped improve VST scores. At the same time, however, the researchers stated that there were a few issues to consider. First, the control group's low scores could have resulted from the fact that the researchers had not taught that group themselves, and apparently, participants had not spent strictly two hours weekly on the assigned readers. Secondly, the researchers further explained that the difference between the experimental groups could have been caused by the fact that the vocabulary group had shown a higher level of English prior to the study. Therefore, WordEngine had probably automatically assigned less frequent words for them to work on. Those words had a lower possibility of encounter in class, which could have led to lower exposure, and eventually less learning. The authors finally concluded that the vocabulary group had probably lost motivation being forced to work for two hours on the platform. This could have explained why the vocabulary group scored lower in the post-test than the vocabulary + ER group.

#### *2.2.3.3.2 Commentary*

McLean et al. (2013) seems to be a very particular study since the researchers left important variables to be decided by the flashcard platform itself. The system automatically decided the target words to use, and the spaced repetition schedule. The peculiar side of the study is that the researchers never knew exactly what words each student was actually learning and also, they could not know for sure the actual words participants had finally acquired. The interesting factor of the study is that there were positive results even when there was no strict control over important variables. The introduction of the automatic flashcard program apparently helped obtain larger vocabulary gains quicker than the implicit method alone, and it also seemed to help improve external tests results. This agrees with previous findings (e.g., Nation, 2007; Fitzpatrick et al., 2008; Schmitt, 2008) stating that explicit learning can help learn larger amounts of vocabulary faster than implicit learning.

There are a few shortcomings that arise from this study, however. First, the platform uses an arbitrary interstudy interval system, which is problematic because it is not clear what retention interval it was based on. As research has shown before (e.g., Bahrick, 1979; Cepeda et al., 2006; Cepeda et al., 2008), the retention interval should be decided first before establishing the ISI lags. Second, the final results in VST scores cannot be considered as resulting directly and uniquely from the intervention. Subjects were taking two English classes a week as part of their regular language course, but only one (productive) was used as part of the project. The receptive class could have also contributed to the overall vocabulary gain. Another limitation of the study could have been the fact that all three groups worked on different in-class activities after reading the books which could have also influenced vocabulary acquisition and final score results.

To conclude, although a few questions arise regarding the validity of the findings, McLean et al. (2013) still managed to blend several topics related to spaced repetition in one longitudinal study. The authors seem to have succeeded in integrating an online flashcard program for a whole year into a set curriculum. Participants overall seemed to respond well to the intervention and to the combination of explicit and

implicit teaching methods. In a very similar fashion, Milliner (2013), reviewed below, also tested vocabulary acquisition introducing an online flashcard program. The researcher in the study aimed at testing the validity of the study by means of an external English skills test at the end, that could also contribute to keep learners engaged and motivated.

#### **2.2.3.4 Milliner (2013)**

This study by Milliner resembles McLean et al. (2013) in that it used Japanese university students as participants and an online flashcard program was employed to improve vocabulary learning. Both studies also included incidental learning, and participants took the Vocabulary Size Test (VST) (Nation & Beglar, 2007) to check vocabulary gain. Milliner (2013), however, expanded McLean et al. (2013) by also introducing an external skills test (TOEIC) to check learning. The Test of English for International Communication (TOEIC) is a standard test to check every day English skills of people in international environments. TOEIC preparation materials were also used in the learning sessions as incidental learning, and to also teach participants exam-specific skills to sit the TOEIC test at the end of the intervention.

Based on the ideas of several scholars (e.g., Schmitt, 2008; Spiri, 2008; Elgort, 2011; Nakata, 2011, McLean et al., 2013) Milliner decided to investigate the effectiveness of explicit vocabulary learning as a robust technique to quickly learn large amounts of vocabulary that could also enhance retention of knowledge. Therefore, the author conducted a longitudinal study during a full semester in order to increase participants' vocabulary sizes and improve their TOEIC scores. This project also served as a pilot study to test whether an online vocabulary learning platform could be incorporated fully into an actual university L2 English course.

##### **2.2.3.4.1 Summary**

The researcher introduced Quizlet ([www.quizlet.com](http://www.quizlet.com)) as an online explicit learning platform, thinking that digital flashcard programs can offer more opportunities for learning than flashcards on paper. The author further explained that students can be more creative with online platforms (applying formatting options and by adding multimedia) and are more engaged in their learning.

In order to better prepare participants for the TOEIC exam, a specialized TOEIC course book was used in class to work on exam specific skills. In order to enhance incidental exposure to the target words, graded readers were assigned weekly, together with a report on them due at the end of every week. For explicit vocabulary learning the researcher prepared bilingual (English - Japanese) weekly sets of 100 flashcards each.

The subjects in the study were 42 Business Management students at a Japanese university taking English as L2. The students were highly motivated to learn since they wanted to score high in the TOEIC test. This high motivation was based on the fact that some of the students were planning on applying to overseas exchange programs, also because they needed good grades to graduate at the university, and because high TOEIC scores would help them obtain a good job after graduation.

In order to gauge subject learning, participants took the VST test at the beginning and at the end of the intervention, a TOEIC test prior to the beginning of the project, and another one before the end of the study. To obtain subject impressions on the intervention, participants' reflections on the study were collected at the end through a post-intervention questionnaire.

The intervention consisted of 30 lessons of 100 minutes each, twice a week, for a duration of 15 weeks (a full semester). In the first lesson participants were introduced to Quizlet as a tool to help learn vocabulary explicitly for the TOEIC test. In the second lesson students took the VST test to check receptive vocabulary knowledge and to be used as a pre-test for base data. In the third session students familiarized themselves with Quizlet and created flashcards of their own. Familiarizing and set-up lessons continued until week four. Starting from week four (until week ten), participants received a set of 100 bilingual flashcards per week corresponding to vocabulary taken from one chapter at a time from the course textbook. During those lessons participants worked about 15 minutes on Quizlet, and then they switched to other TOEIC related activities. Participants were also assigned a graded reader a week and had to write a report on it by the end of every week. At several points during the



intervention the teachers directed subjects to take quizzes generated by the system enforcing this way repeated exposures to the vocabulary items.

On the 12<sup>th</sup> week of the intervention participants took the TOEIC test. On the last learning session they took the VST post-test and they completed the questionnaire to reflect on the intervention.

In order to analyze data, TOEIC results from scores prior to the intervention and test scores from week 12 of the intervention were considered. VST scores were obtained from tests taken during the second session and the last session of the intervention. Table 2-10 below shows average scores for all participants for all tests, and the corresponding increase in percentages for each test type. Results clearly show that, as a whole, participants scored higher in the two tests after the intervention suggesting that the use of the flashcard platform probably helped improve their test scores.

<b>Test</b>	<b>Before intervention</b>	<b>After intervention</b>	<b>% Increase</b>
TOEIC	411	464	12.9%
VST	6454	6956	7.78%

**Table 2-10: Percentage scores comparison of TOEIC and VST tests (Milliner, 2013, p. 56)**

Subject questionnaire and survey results revealed that 68% of the participants agreed that the digital platform was a good way to study for the TOEIC exam, and that they enjoyed working with the software.

The researcher concluded that the intervention seemed to have contributed successfully to increase scores results of both external tests taken by the subjects. The addition of the digital platform for explicit learning apparently contributed to better learning and to higher participant engagement and enjoyment.

#### *2.2.3.4.2 Commentary*

Milliner (2013) succeeded in blending explicit learning through an online platform with a set curriculum to prepare students for a final external exam (TOEIC). The researcher of the study obtained the expected results and participants were motivated to work and enjoyed the use of Quizlet for vocabulary learning. As opposed to McLean et al. (2013) participants in Milliner (2013) were highly motivated, in part probably due to the fact that they saw an actual use (improving TOEIC scores) of the intervention requirements.

This study concentrated more on learning per se than on long term retention. Nevertheless, teachers leading the lessons instructed subjects to take system-generated quizzes on the target words, as a way to introduce spaced repetition to the sessions. Since there was no retention interval or interstudy interval systematically planned it is difficult to analyze whether the quizzes on Quizlet helped with vocabulary acquisition after all. At the same time, in the study, Milliner seemed to have erroneously assumed that the sole inclusion of explicit vocabulary learning increases vocabulary retention over time. This is against a long line of research (e.g., Bahrick, 1979, Bloom & Shuell, 1981; Cepeda et al., 2008) claiming that long-term retention is not based on the learning methodology itself (for instance, explicit or implicit learning), but rather on the repetition of information based on a proper integration of the interstudy interval and the optimal combination between the retention interval and the interstudy interval.

At the same time, VST and TOEIC results cannot be taken as solid indicators of correlation between the Quizlet integration and test scores in this study. It could be argued that even without the inclusion of the flashcard platform, participants could have still improved their test scores just by following regular course requirements. The inclusion of a control group (following course requirements but not working on Quizlet), for instance, could have helped measure the actual dimension of the intervention by contrasting pre-test and post-test scores of an experimental group against a control group.

It can be concluded that the introduction of the TOEIC final exam added an extra component to the intervention. This probably contributed greatly to motivate participants that learned the language with a particular goal in mind (succeed in the TOEIC test). Motivation and engagement are of particular importance in longitudinal studies as participants might lose motivation after working on the project for some time, and therefore, not commit fully to the study requirements. Also, the lack of data about the target words that participants had finally acquired, and the unplanned inclusion of spaced repetition (as seen in this study and also in McLean et al., 2013) seem to stress the need to conduct research studies investigating actual learning and retention of target words. Therefore, in a spaced repetition longitudinal study to actually test vocabulary acquisition and long-term retention, it seems more logical to teach participants a set of target words, carefully plan retention intervals and interstudy intervals, monitor participant vocabulary learning, and later check retention levels.

Finally, the three articles reviewed so far in this subsection concentrated on innovative teaching methods to improve acquisition and retention of vocabulary. Erbes et al., (2010) found that the implementation of a new methodology with younger learners did not show any retention gains after all. On the other hand, McLean et al., (2013) and Milliner (2013) obtained satisfactory results with the retention scores using young adults as participants. Considering the age of participants in Erbes et al., (2010) was different from the other two studies, it is still unclear whether the success of innovative methods in retention scores could depend on the age of the participants. Gryzelius (2016) tackled this issue by conducting a field study introducing an online flashcard program to improve retention in an L2 Spanish course using middle-school learners as subjects. This study is reviewed next.

#### **2.2.3.5 Gryzelius, (2016)**

Gryzelius saw that language students in Swedish schools managed to learn vocabulary properly, but the new words did not stay for long in students' memories. In general, those students tended to forget most of the vocabulary after they took their final exams.

Therefore, similar to the other three articles in this subsection, Gryzelius (2016) also introduced an 'innovative' method. The researcher of the study implemented spaced repetition (SR) into an L2 Spanish classroom as opposed to the everyday (traditional) teaching methods the students would normally be confronted to.

The researcher wanted to investigate whether spaced repetition could help L2 Spanish young learners in a Swedish school improve their vocabulary retention over time. In order to improve vocabulary learning, the researcher opted to teach vocabulary explicitly through the use of an online flashcard program. The researcher also decided to let participants learn independently, because he was more interested in the spacing effect rather than in the way students studied.

#### ***2.2.3.5.1 Summary***

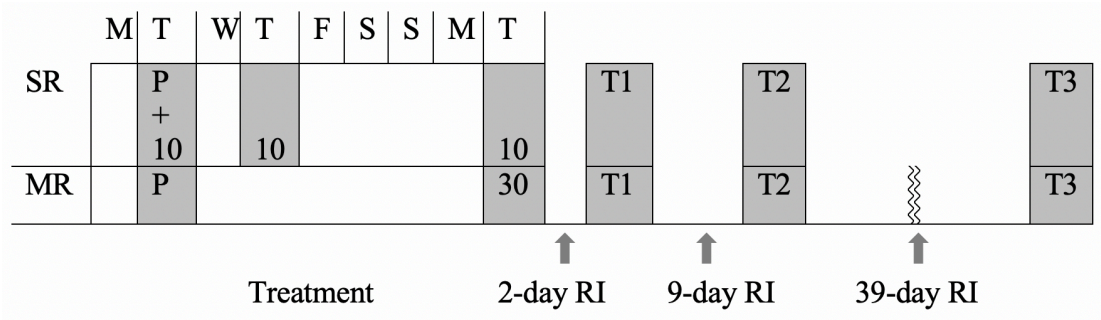
To improve vocabulary learning the researcher decided to use Glosor.eu, which is an online learning platform. The system allows students to create their own digital flashcards and it also contains games and several entertaining activities to enhance learning. Similar to McLean et al. (2013) and Milliner (2013), Gryzelius (2016) also employed an online platform (rather than having participants working on paper) with the idea that participants would be more entertained and motivated. According to Gryzelius (2016) participants in the study were already familiar with the digital flashcard platform which helped them with engagement and to work more confidently.

Gryzelius (2016) replicated some of Bloom & Shuell's (1981) ideas by using 20 target words and eliminating cognates. Participants were expected to learn 20 target words selected from the course textbook, from chapters students had not covered in class yet. The total number of target words (20) was based on the fact that subjects were already used to learning 20 new words per week in their regular course of studies.

Participants were 30 students in grade eighth in a Swedish school taking L2 Spanish. The main group was divided into two conditions, massed (M1) and spaced (notice that the article refers to spaced repetition as distributed practice) (D1), with 15

students each. For consistency across this thesis, only massed repetition (MR) and spaced repetition (SR) will be used.

The project started with an introduction to the target words, plus 30 minutes of learning, followed by three post-tests. Participants were divided into two groups and were trained according to the condition they were in (massed repetition or spaced repetition). The first day both groups were introduced to the target words as word pairs in order to learn meaning, pronunciation and spelling of the words. The target words were written on the board by the researcher. After the presentation, the spaced repetition group continued to work on the target words for ten minutes. They started to work on Glosor.eu for eight minutes on a multiple-choice activity. For the remaining two minutes of the session participants worked on a translation activity. The group practiced again on the target words for ten minutes two days after the first session, and again five days after that. In those two sessions participants worked on the same activities as before, and they were also allowed to play the available games on the platform. There was a post-test one day after the last learning session, a second test seven days after the first test, and a third test thirty days after test two. The researcher decided to use an expanding interstudy interval (ISI) to match the days when participants had their usual Spanish class at school. Also, the author claimed that information needs to be repeated more quickly at the beginning of the learning phase, but after information has been acquired it is better to expand the lag between learning sessions. The massed repetition (MR) group had the presentation of the target words on the same day as the other group, and then studied the target words for 30 minutes on the same day the spaced repetition (SR) group had the last learning session. The massed repetition (MR) group had the same test schedule as the other group. Figure 2-9 below shows the project schedule where the P represents the presentation session of the target words, and 10 and 30 show the number of minutes spent rehearsing the words. The retention interval (RI) number of days were counted starting from the last learning session. T1, T2, and T3 represent each one of the post-tests in the study.



**Figure 2-9: Project schedule (Gryzelius, 2016; p. 12)**

All three tests were identical and consisted of a translation activity on paper where students were asked to provide the Spanish translation of a given Swedish word.

Since several students were absent during the days when the project took place, only the ones present in very session and in every test were considered for the results. As a consequence, only 21 participants overall were included in the results. This resulted in eight final subjects in the spaced repetition group, and 13 in the massed repetition group. Table 2-11 below shows mean values of all tests for each group.

Group	Test		
	T1 (2-day RI)	T2 (9-day RI)	T3 (39-day RI)
SR	93.12%	94.35%	85%
MR	50%	51.15%	46.15%

**Table 2-11: Mean percentage of correct responses of the three tests (Gryzelius, 2016, p. 16).**

Results above show that the spaced repetition condition obtained higher results in all three tests, revealing that spaced repetition was a more successful method for retention of information than massed repetition in the current study. Gryzelius stated that the introduction of spaced repetition was successful as 30 days later participants were still able to remember 85% of the target words. Despite the fact that the author

obtained expected results, he still reported lack of motivation and commitment by some of students.

#### **2.2.3.5.2 *Commentary***

Gryzelius's (2016) combination of strategies to teach L2 Spanish vocabulary to middle-school students demonstrated that spaced repetition can be beneficial for long-term retention of information. The fact that the intervention took place during regular class hours, and that participants were allowed to study independently on a platform they already knew, added certain easiness to the field study and probably contributed to the spaced repetition gains of the project.

Although spaced repetition findings in this study are conclusive, this cannot necessarily be attributed to the expanding lengths of the interstudy intervals (ISI) used in the study. The author stated that he had purposely selected an expanding interstudy interval as it can contribute to better retention than an interval that is always of the same length. However, there is no clear evidence regarding the benefits of using expanding or fixed-time ISIs. For instance, researchers (e.g., Cepeda et al., 2006; Kang et al., 2014) did not find any particular difference in retention gains for either option (expanding or fixed interstudy intervals) and Nakata (2015) only found a limited advantage of expanding over fixed intervals.

To conclude, Gryzelius (2016) implemented spaced repetition in an L2 Spanish course and its findings reveal that the introduction of the new methodology to an everyday classroom can also be successfully applied to younger learners. Similar to McLean et al., (2013) and Milliner (2013) the online flashcard program apparently promoted learning and retention. Finally, what is particularly intriguing, despite the rather confounding interstudy intervals used in the study, is that when analyzing Gryzelius' (2016) results (see Table 2-11), retention findings concur with Bahrick's (1979) results in that the retention curve does not drop after rehearsals, but instead, it continues to rise before starting to drop.

### **2.2.3.6 Conclusion**

The four articles in this section showed that innovative teaching methods looking to improve retention of information can be applied to real classroom settings but there are some important considerations to be made beforehand. As Erbes et al., (2010) demonstrated, careful planning is needed to avoid having external variables (e.g., student teacher relationship, motivation) negatively affecting study results. Secondly, blending everyday teaching methods with an online flashcard platform for explicit vocabulary learning seemed to work well with different age groups and to promote quick learning and retention as shown in McLean et al. (2013), Milliner (2013) and Gryzelius (2016). Finally, the inclusion of a formal final assessment (e.g., TOEIC in Milliner, 2013) seemed to motivate learners further and probably also contributed to the study's positive results.

In a rather similar fashion, the next subsection will refer to a formal external component that L2 learners sometimes are confronted to. As it is the case in IB Ab Initio courses, learners are confronted with two graded assignments with authentic materials at the end of the course. This can be overwhelming for beginner students, but preparing them to tackle authentic materials, and by including these materials to a language course could help them with the assignments and could also help with motivation. In this sense is that Johnson & Heffernan (2006) introduced spaced repetition to promote learning, retention and comprehension of authentic materials. The article is reviewed next.

## **2.2.4 SR to help with authentic materials, motivation, and retention**

### **2.2.4.1 Introduction**

This subsection will review a very particular study (Johnson & Heffernan, 2006) because it covers several of the topics referred to in the introduction of this chapter (e.g., vocabulary learning, spaced repetition, long-term retention, motivation, comprehension of authentic materials). Johnson & Heffernan (2006) is also at the same time the only study that directly focuses on non-advanced L2 learners and their difficulties to comprehend authentic materials. The article also examines spaced



repetition as a method to enhance long-term vocabulary retention, while also focusing on participant motivation and engagement.

Two of the articles reviewed above (McLean et al., 2013 and Milliner, 2013) investigated how, after the intervention, participants performed in tasks (external exams) that were not especially prepared for the project (e.g., VST and TOEIC). In a rather similar fashion, Johnson & Heffernan (2006) introduced authentic materials that were not especially made for the study. Authentic materials are very problematic for L2 beginner learners since their vocabulary knowledge is very limited. In order to comprehend authentic texts large amounts of vocabulary are needed (Hu & Nation, 2000; Nation, 2006; Schmitt, 2008) therefore, beginner learners tend to struggle when they are confronted with them. Johnson & Heffernan (2006) is the only study in this literature review that directly referred to authentic materials as a very special topic on its own. The article investigated whether participants' comprehension of authentic materials was affected as a consequence of the intervention. This article also deserves special attention since the main study of this thesis prepares subjects (L2 Spanish beginners) to eventually work on two final tasks containing authentic materials required to pass the course.

Together with the fact that Johnson & Heffernan (2006) tackles authentic materials it also appears as a very attractive study to replicate considering that it concentrates on motivating participants as well. Motivation certainly deserves a special mention as some of the scholars reviewed above reported lack of engagement and commitment (e.g., Erbes et al., 2010; Goossens et al., 2012; and Gryzelius, 2016). Johnson & Heffernan (2006) employed movie trailers to motivate subjects. This seems to be a good strategy considering that videos have been reported to be a very good source of authentic materials, and also as a very efficient way to motivate and engage learners (e.g., King, 2002; Sherman, 2003; Zanón, 2006).

Finally, Johnson & Heffernan (2006) also made use of spaced repetition aiming at long-term retention of target words. Arguably, however, the methodology employed in the study appeared to have erroneously tackled the issue. This will be discussed later in the commentary section of the article review.

#### **2.2.4.2 Johnson & Heffernan (2006)**

Johnson & Heffernan conducted a study focusing on repeated exposures and innovative methods in L2 vocabulary learning to improve language comprehension and motivate learners. The authors investigated the assumption that a recycling method can help with vocabulary acquisition and retention over time.

The authors hypothesized that students' knowledge of target words can be enhanced through a high concentration of repeated exposure of vocabulary in written and audio-visual media in a confidence building learning experience. Therefore, the authors created a movie-trailer project based on spaced repetition with the idea of enhancing long-term retention.

This article focused on authentic materials because Johnson & Heffernan believed that foreign L2 learners at the beginner level struggle with authentic materials when they progress to higher levels. Therefore, the researchers created a project claiming that when L2 learners are able to understand authentic materials, their confidence is boosted and their desire to continue studying the target language is prolonged. The authors conceived this project with the notion that in most formal instructional settings, language learners are taught using graded materials, and are introduced to real life resources once they reach more advanced language levels.

##### **2.2.4.2.1 Summary**

The study consisted of a series of 15 short readings, where 112 target English words were taught and assessed in order to assist students in comprehending ten movie trailers. The trailers contained multiple high-frequency words and the target words were repeated across the 15 different short readings. After reading all 15 texts, students have been exposed to all target items at least three times.

In order to help participants learn the new words, target vocabulary items in the short readings were clickable, providing part of speech, definition, example and, when applicable, a picture. Whenever possible, contextual clues were placed around target items in the readings to help guessing meaning from context. After each reading, students were quizzed on their comprehension of the target vocabulary. Target items

were further reinforced by the recycling method when students watched the trailers later.

The participants were 119 first and second-year Japanese university students. The subjects were taking mandatory English courses required for graduation.

The whole project took place over nine weeks. The first week participants took the pre-test. During the following seven weeks subjects did two readings per week. During the ninth week they did one last reading and took a post-test. In order to ensure that participants read and finished every reading activity, each of the 15 short readings was followed by a series of questions. Only after the participants had answered all questions correctly, the trailer for that reading would become available, and participants could watch it. To check learning of target vocabulary, participants took a post-test that was an exact replication of the pre-test at the end of the intervention.

The pre-test and post-test consisted of two sections. Section one included ten context-based questions, similar to the questions following each reading. Each question focused on one target word randomly selected. Section two included multiple-choice questions based on ten short clips from the trailers making use of ten randomly selected target words.

Section	Project tests	
	Pre-test	Post-test
Vocabulary section	56.5 (11.9)	72.5 (11.3)
Video section	56.8 (13.5)	64.9 (16.3)

**Table 2-12: Mean scores of project tests with standard deviation between parenthesis (Johnson & Heffernan, 2006; p. 73)**

Table 2-12 above shows results of all tests with their respective standard deviation. Findings showed that when pre-test and post-test results were contrasted, there was a

16.0% increase in scores in the vocabulary section, and 8.1% in the video section. The researchers expected this difference between sections, since questions in section one were similar to the ones met after the readings. Although when contrasting the pre- and post-test participants made significant gains, they were not as high as Johnson & Heffernan had initially expected. The researchers further explained that results were different from what they had anticipated since student effort and commitment was at times questionable.

To sum up, despite the fact that retention gains were lower than originally expected, the authors concluded that the initial objective of creating an online activity to boost learners' comprehension of authentic materials had been met.

#### ***2.2.4.2.2 Commentary***

The authors introduced a very interesting concept of enhancing comprehension of authentic materials while also aiming at participant motivation. The movie trailers added a quota of authenticity to the project, while recycling vocabulary through repeated exposures seemed to have enhanced learning and comprehension (as inferred from test results). At the same time, the combination of learning implicitly (through the readings and videos), and explicitly (through the use of the hyperlinks offering the meanings of the target words) provided an opportunity for participants to learn independently. In the readings, participants had the opportunity to check the meaning of the target words or inferring them from context. This method most probably helped build comprehension strategies using contextual clues needed for comprehension of authentic materials. This method of using contextual clues is a well-documented strategy used by learners as they tend to rely heavily on the surrounding context to comprehend unknown words (Parry, 1991). The main project of this thesis could benefit greatly from such a methodology to assist subjects in comprehending authentic materials as well.

Johnson & Heffernan (2006) appears to have a few limitations. Firstly, although the study aimed at long-term retention, it never really tested it. A post-test after the learning sessions to check retention of target vocabulary could have helped the researchers obtain data to analyze whether information was remembered. Secondly,

the authors of the study mistakenly assumed that by repeated exposures alone vocabulary would be retained longer. Although information is retained longer when repeated, as opposed to not repeating it, as it has been documented before (e.g., Dempster, 1988; Schmidt & Bjork, 1992; Bahrick, 2005), there should be careful planning of the precise times when information needs to be repeated. As stated by Bahrick (1979) and Fitzpatrick et al. (2008) more rehearsals contribute to better learning and retention is enhanced, however, how long information is retained depends heavily on the combination of the retention interval and the interstudy interval (e.g., Bahrick, 1979; Cepeda et al., 2006; Cepeda et al., 2008). Therefore, leaving repetition to uncertain and unplanned encounters does not seem to be the most effective way of enhancing long-term repetition.

Another issue that arises from the researchers' final conclusions is that they reported lack of student commitment and motivation. Therefore, it seems that one of the main goals of the study could not be met after all (increasing learner motivation through an enjoyable activity). A survey at the end of the project could have helped the authors learn why some participants were not actively engaged in the activities.

To conclude, this project tackled five main issues referred to in the introduction of this chapter: long-term retention, spaced repetition, authentic materials, teaching techniques, and student motivation. Those five issues are crucial aspects to consider in a research project having IB Ab Initio students as participants, as discussed in the Introduction chapter. To start with, IB Ab Initio students have less than two years to learn vocabulary on a large set of different topics. This means that topics learned in the first year will not be revisited the following year, therefore, retention of previously learned knowledge is crucial for success. Secondly, this retention can be enhanced through the implementation of spaced repetition in the course, and the appropriate combination of the retention interval and the interstudy interval. Third, in a personal written task and in the final reading exam students are confronted with authentic materials, which means that there will be a long list of unknown words in the texts. Finally, student motivation tends to diminish when high-school graduation approaches, therefore there is a need for teaching techniques that could improve learning and keep students motivated.

## **2.3 Discussion**

This Discussion will present the final conclusions regarding the articles reviewed across this chapter, in the different sections, highlighting the salient findings and what is still missing regarding spaced repetition and long-term vocabulary retention. The research questions guiding this thesis appear as a consequence of this review and are listed further below.

The conclusions below are presented following the section order of the literature review. However, since all of the articles reviewed in this chapter, while different, still shared some common aspects regarding spaced repetition and retention of information, their findings and methodologies are contrasted against other articles in different sections to provide more conclusive remarks.

### **2.3.1 RI/ISI findings**

To start with, the first section in this literature review exposed the fact that almost 40 years after Bahrick's (1979) findings, conclusions regarding the ideal combination of retention interval (RI) and the interstudy interval (ISI) are still rather unclear. The other aspect that does not seem to be unanimously agreed upon is the notion by Bahrick (1979) and Cepeda et al. (2008), for example, that retention will be at its highest at the optimal RI/ISI combination, but it will be lower before and after it providing an inverted-U curve of retention of information. The last topic that arises after analyzing the articles in the first section is retention gain. This refers to the fact that although information is retained in some cases even after 350 days after the last learning session (e.g., Cepeda et al., 2008), some of the reviewed papers show, arguably, low retention scores. The issue that emerges then, is not only how long information can be retained, but also, how much of it is still remembered.

Although the articles reviewed in the very first section (see 2.1 above) agreed on the fact that the interstudy interval increases as the retention interval increases, they had major differences in relation to the length of the interstudy interval given a certain retention interval. While Bahrick (1979) and Lotfolahi & Salehi (2017) found that the interstudy interval should be equal in length to the retention interval for best long-term retention results, Cepeda et al. (2008) and Küpper-Tetzel et al. (2014) suggested

that the value of the interstudy interval should be a portion of the value of the retention interval, and that the larger the retention interval, the smaller portion of it is needed to represent the ideal value of the interstudy interval. This is also in line with what other researchers have found (e.g., Rohrer & Pashler, 2007; Suzuki & DeKeyser, 2017; Serrano & Huang, 2018). For instance, for receptive knowledge, Cepeda et al. (2008) claimed that the optimal value of the interstudy interval should be around 20% of the retention interval for a 35-day RI, and an interstudy interval of 7% of the retention interval would be ideal in 350-day RI.

The second aspect that this literature review has exposed is the fact that findings from Küpper-Tetzel et al. (2014) are not conclusive as the article shows agreement (based on the 7-day RI) with the inverted-U curve found in Bahrack, (1979) and Cepeda et al. (2008), but results are confounding at 10-day ISI. Even more, results from Fitzpatrick et al., (2008) do not seem to comply with the inverted-U curve at all. This therefore also highlights the need to continue to conduct research on the field that could help understand the RI/ISI combination even further.

Finally, when the main goal is to implement spaced repetition into real educational settings, it is important how long information is retained, but most importantly it is crucial to determine how much of that information is still remembered. For instance, in courses such as IB Ab Initio, where learners are required to sit a final exam (which is of particular importance for passing the course and university admissions) it is imperative that they can retain most of the knowledge previously learned in order to succeed in their tests. Along this line, Table 2-13 below lists studies across the different sections in this literature review and contrasts how much of the information learned is retained after the intervention. Notice that although all of the articles reviewed above were included in the table, some of them did not show exactly how much information was retained as a result of the intervention (e.g., Johnson & Heffernan, 2006; McLean, 2013; and Milliner, 2013).

Considering the fact that this thesis proposes a longitudinal spaced repetition study, it seemed more relevant to include in Table 2-13 only the longest RI/ISI combination in each article (in case they included more than one RI/ISI condition) to analyze

retention gains in the longest possible retention interval. At the same time, in the column Extension of study only the learning section of the intervention without the retention interval period was considered. Also, only productive knowledge results were included of studies that tested both (productive and receptive) since all of the studies listed tested productive knowledge, but not all them tested receptive knowledge. Only receptive results of Küpper-Tetzel et al. (2014) were considered, however, since productive results provided in the article had undergone some statistical analysis and were not comparable to those of the rest of articles in the table (which were raw results). Also, since in Erbes et al. (2010) there are multiple inconsistent values only the highest of the scores at 24-day RI were selected, and in Gryzelius (2016) the largest interstudy interval was selected to focus only on the largest portion of the learning lag. Finally, the Age column adds extra information to the table regarding the age of the subjects in the study. Uni refers to university students, PS, MS, and HS refer to primary, middle and high-school in the American system, and Adult and Mixed refer to adult subjects (non-university students), and mixed ages respectively.



Article	Age	Lab /field	Extension	Knowledge	Sessions	ISI (days)	RI (days)	%
Bahrnick (1979)	Uni	L	6 months	P	6	30	30	95%
Bloom & Shuell (1981)	HS	F	3 days	P	3	1	4	75.2%
Johnson & Heffernan (2006)	Uni	L	9 weeks	R	15	2-5	-	-
Fitzpatrick et al. (2008)	Adult	L	20 days	P	20	1	42	45%
Cepeda et al. (2008)	Mixed	L	42 days	P	2	21	350	19%
Erbes et al. (2010)	MS	F	1 day	P	1	-	24	37%*
Sobel et al. (2011)	MS	F	14 days	P	2	7	35	20.8%
Goossens et al. (2012)	PS	F	4 days	P	4	1	35	49.49%
McLean (2013)	Uni	F	28 weeks	P/R	28	7	-	-
Milliner (2013)	Uni	F	15 weeks	P/R	30	2-5	-	-
Küpper-Tetzel et al. (2014)	MS	F	20 days	R	2	10	35	54%
Gryzelius (2016)	MS	F	8 days	P	3	5*	39	85%
Lotfolahi & Salehi (2017)	PS	F	14 days	P	2	7	35	50.71%

**Table 2-13: Comparison of retention gain showing percentage means**

Table 2-13 above demonstrates that the general retention gain average (considering the percentages reported of all the articles listed in the table combined) was 54.91%. This overall retention mean seems to be rather low if the goal is to obtain the maximum possible retention gain for learners to, for example, succeed in final exams. The highest gain in the table is shown by Bahrck (1979) in a lab study. It could be arguably stated that a retention gain of 95% (of the target words) is sufficient for learners to do well in final tests. This study by Bahrck is at the same time the longest of them all with a learning phase extending for six months. The second highest scorer on the list is Gryzelius (2016), which is a field study, but very short in duration (8 days). The longest field study (the learning phase extended for 20 days) listed on the table is Küpper-Tetzel et al. (2014), but retention results seem to be low as well: 54%. This suggests that there is also a need for field studies extending over months that could provide comparable retention gains to those of Bahrck's (1979). This very last statement is actually controversial and double folded. A test where most participants obtain such a high score runs the risk of producing a ceiling effect. The ceiling effect is described as the effect in which participants reach the highest maximum score in a study simply because the test was not difficult enough (Ary et al., 2018). While retaining as much as possible would be the main goal of an actual language program so learners can do well in tests, it is conflicting in language research since it impedes participants from learning more. Therefore, further research on long-term vocabulary retention should have this in mind in order to investigate strategies to enhance high retention gains, but without running the risk of producing a ceiling effect.

### **2.3.2 Learning findings**

The second section of this chapter focused on the teaching and learning portion of the articles and methodologies employed. The section also provided very interesting findings and it exposed two major gaps (described below) in the spaced repetition and long-term retention field. Findings from the different subsections that could inform the planning of the main project are described below, followed by gaps that the same project will try to fill.

To start with, findings from Fitzpatrick et al. (2008) demonstrated that learning vocabulary explicitly, combined with a system of learning and rehearsal (which is in

line with Bahrick, 1979) can help acquire large amounts of vocabulary quickly. The same article reported that knowledge of target words decayed quickly the moment rehearsals had stopped. Secondly, the three following articles analyzed (Bloom & Shuell, 1981; Sobel et al., 2011; Goossens et al., 2012) bridged the gap between lab and ecological studies and showed that spaced repetition could be applied to field studies, successfully, seamlessly and across age groups. Those three studies however, were very short in duration, which emphasizes the need to see whether those findings still hold in longer studies. The four articles reviewed afterwards seemed to enhance learning and retention through innovative teaching methods. As seen in Erbes et al., (2010), the implementation of a new method should still be closely monitored and carefully planned in order to prevent unexpected results. Lastly, Johnson & Heffernan (2006) received special attention considering it touched upon topics covered in the previous subsections and it aimed at helping beginner L2 learners comprehend authentic materials. This is a very interesting project to replicate considering the main project of this thesis shares similar objectives to those of Johnson & Heffernan's (2006), and because a replication study could also be used as a pilot study to pre-test logistics and avoid deficiencies in the main project of this thesis.

One of the first gaps found across this literature review is the fact that there is a lack of longitudinal field studies (extending over several months) having high-schoolers as participants. For instance, the three longest studies reviewed above had university students as participants (Bahrick, 1979; McLean et al., 2013; Milliner, 2013), and the longest one having high-schoolers as subjects was Küpper-Tetzel et al. (2014) with a learning phase of only 20 days. This therefore highlights the fact that to comprehend whether spaced repetition enhances long-term vocabulary retention in high-school students, and how spaced repetition can be successfully implemented into a high-school environment, more longitudinal field studies are needed with this age group.

The second gap refers to the fact that the field studies reviewed focused too much on the difference between spaced and massed repetition, but not so much on the best way to implement spaced repetition in the curriculum. For example, most of the studies mentioned above (e.g., Bloom & Shuell, 1981; Sobel et al., 2011; Gryzelius, 2016; Lotfolahi & Salehi, 2017) claim that spaced repetition improves long term retention.

Considering those studies reported a lack of spaced repetition integration into school curriculums, and stated that more field studies are needed, then future research should conduct field studies aiming at the best implementation of spaced repetition in real educational settings, instead of just contrasting the benefits of spaced vs massed repetition, as seen in studies listed above. Therefore, the question should not really be whether spaced or massed repetition brings major gains with long-term retention, but rather, it should investigate the ideal way of introducing spaced repetition so larger portions of knowledge are remembered, and for longer periods of time.

To conclude, if spaced repetition plays a significant role in long-term retention of information then knowing exactly when to repeat is crucial to ensure that information is not forgotten. Researchers who have studied spaced repetition and retention of information argue that given a certain retention interval (RI), there is an optional gap between learning sessions (interstudy interval) that will produce higher retention results. However, scholars still do not seem to agree on the optimal combination between the retention interval and the interstudy interval (ISI). At the same time, most of those studies have been conducted in laboratories. There seems to be a rather small amount of field studies, but they tend to be short in duration. Those studies, at the same time, have mostly had university students as subjects, and have focused specially on the retention gains of spaced repetition against massed repetition (rather than contrasting spaced repetition against an actual course following a traditional teaching method). Finally, most of the previous research studies seemed to concentrate on the length of time that information is being retained in the brain, but not so much on how much of that information is kept. As a consequence, this dissertation will focus on whether spaced repetition can enhance L2 vocabulary acquisition and long-term retention in a field study having high-schoolers as subjects. This thesis will also try to provide further data regarding the optimal RI/ISI combination.

### **2.3.3 Research Questions**

In response to the claim that more longitudinal field studies with young learners are needed, this thesis proposes a field study where conditions are kept as ecological as possible. In order to test whether high-school seniors can actually benefit from the

introduction of spaced repetition in a language course in comparison to a traditional teaching method (the one used by teachers every day in the classroom), the following questions arise.

1. Will spaced repetition produce different retention levels in participants in the experimental group in comparison to students who were taught using traditional teaching methods and graduated a year earlier?
2. Will spaced repetition produce different retention levels in comparison to a control group who were taught using traditional teaching methods?

As discussed above, it is still rather unclear which best combination between the retention interval (RI) and the interstudy interval (ISI) can produce better retention levels. Therefore, by repeating information every 30 days, the following question will be addressed.

3. Given a 30-day ISI, which RI provides the highest retention scores: 30-day, 60-day, or 70-day RI?

## **Chapter 3: Replication of Johnson & Heffernan (2006)**

### **3.1 Introduction**

A replication study repeats a previous study in such a way as to extend, limit, or reconsider previous findings. The main objective of the replication study is to investigate whether previous findings are reliable and/or can be generalized to other contexts (Porte, 2012). This chapter introduces an approximate replication study in which major variables (from the original study) remained unchanged in order to compare findings between the original and the replication study. The main goal of this replication is to investigate whether findings from Johnson & Heffernan (2006) are generalizable to high-school learners studying L2 Spanish at an international school.

It seemed appropriate to replicate Johnson & Heffernan (2006) since the article contributed to the field by investigating how spaced repetition (SR) could enhance vocabulary retention while increasing understanding of authentic materials at the same time. The scholars also introduced an innovative teaching method using movie trailers aiming at higher participant motivation and engagement.

The study of Johnson & Heffernan is particularly relevant for this thesis since many features in the original study still remain valid for today's vocabulary acquisition. To start with, students still seem to forget previously learned vocabulary quickly, and despite the fact that spaced repetition has been long regarded as a means to enhance long term retention (e.g., Bahrick & Phelps, 1987; Dempster, 1989; Cepeda et al., 2006; Ebbinghaus, 2013), it still has not been implemented widely in educational institutions. At the same time, more spaced repetition field studies seem to be needed to fully understand how retention of vocabulary actually works. Finally, non-advanced learners still seem to continue to struggle to comprehend authentic materials and (as it was seen in some articles in the Literature Review) lack of motivation can negatively affect learning (e.g., Erbes et al., 2010; Goossens et al., 2012; and Gryzelius, 2016).

The present replication investigates whether high-school students taking Spanish as a foreign language improve vocabulary acquisition and retention through Johnson &

Heffernan's movie project, and whether participants' confidence towards authentic materials is boosted. At the same time, this replication also serves as a pilot study to test how memory research could be applied to the actual school where the main project of this thesis will take place.

### **3.2 Summary of the original study**

Johnson & Heffernan saw that in most formal instructional settings, language learners are taught using graded materials, and are only introduced to authentic materials once they progress to advanced levels. When these learners are first confronted with those authentic materials, they feel frustrated and confused. The researchers realized that the introduction of authentic materials at an early stage in language learning, could boost learners' confidence and prolong their desire to continue to learn the target language. Hence, Johnson & Heffernan designed the Short Readings project in order to assist L2 English learners understand ten authentic movie trailers found on a movie web site.

The project consisted of a CALL activity aimed at pre-teaching target items that would later be found in movie trailers. The authors believed that repeated exposures of target items would help subjects comprehend movie trailers and enhance long-term retention of vocabulary. Therefore, the researchers created a series of readings containing those target items to teach the words while also using spaced repetition aiming at improved comprehension and long-term retention of the items.

The target items were 112 English words that were used by the researchers to create a series of 15 short readings to help comprehend ten movie trailers. The participants were 119 first- and second-year Japanese university students taking mandatory English courses as part of their studies. The research project, which extended over nine weeks, started with a pre-test to check knowledge of the target words. This was followed by a series of 15 short readings. Each reading was followed by a series of questions. Once the participants had answered all questions correctly, the trailer for that reading would become available, and participants were able to watch it. At the end of the intervention participants took a post-test that was similar to the pre-test to check learning of target vocabulary. There were two test sections in the original study.

In section one there were ten context-based questions (similar to the ones after each reading) focusing on one target word randomly selected. In section two, there were multiple-choice questions based on ten short clips from the trailers.

Although test results showed that participants had improved their knowledge of target vocabulary, scores were not as high as the researchers had originally expected. For example, in the vocabulary section, group results showed group average results of 56.5% in the pre-test, and this increased up to 72.5% in the post-test. In the video section, on the other hand, the group average results in the pre-test were 56.8% and 64.9% in the post-test. This showed that after the treatment there was a 16% increase in average scores in the vocabulary section, and 8.1% in the video section. The authors attributed this apparent small gain to the fact that some participants were not willing to do their best in the project.

Johnson & Heffernan (2006) seemed to succeed recycling vocabulary since this appeared to enhance learning and comprehension of target items. Despite the fact that the study aimed at enhanced retention, participants were never really tested after the treatment. A posttest after the last learning session could have provided relevant data regarding vocabulary retention. Finally, the researchers of the study believed that participants would enjoy working on the project. However, this was probably not the case considering the researchers reported that participants commitment was questionable at times.

### **3.3 Differences between the original vs. the replication study**

In Johnson & Heffernan (2006) Japanese university students taking English as L2 acted as participants. In this replication study participants were high-school students taking Spanish as a foreign language at an international school in Qatar. In the replication, participant age ranged from 14 to 18, there were 23 different nationalities, and there were several different mother tongues. The great majority of the participants were international students who had not been in the country for more than three years, and for the most part, they already spoke their mother tongue and English (which was the common language of communication and instruction at the school).



Another difference between the original and the replication study was data collection. The original study tested participants prior and after the treatment on their knowledge of the target words. The pre- and post-test were separated in two sections (context-based questions similar to the ones after each reading, and multiple-choice questions based on ten short clips). The replication study in contrast, had three different type of tests, a confidence test, a performance test, and a retention test seven days after the last learning session. The confidence test (see Appendix I) elicited information regarding participant confidence levels before and after the intervention. The performance test (see Appendix II) assessed meaning of target words mixing context-based questions and questions based on seven short clips. The retention test (which was the exact same test as the pre-test) checked how much vocabulary was maintained following the intervention after vocabulary rehearsals had stopped.

Finally, there are two other major differences to also take into account. The original study took place over nine weeks, with one weekly session. The replication on the other hand, took place in three sessions spread across seven days overall. The replication resulted in a seven-day study considering the fact that the school administration specially requested that students were not deviated from their regular curriculum for over a week. Also, although some of the class teachers leading the project lessons were very supportive of the project, they requested that the project did not extend more than three lessons. The final difference to consider between the two studies, is the number and length of the readings, and the number of movie trailers used. There were 15 readings in the original study of 220 to 410 words in length based on ten movie trailers. The replication study, on the other hand, had seven readings of approximately 150 words in length based on seven movie trailers. Shorter readings were used as they resembled the length of the readings participants were used to work with in their regular classes.

### **3.4 Intervention**

This section describes, in detail, the research design of the replication and its implementation. Information regarding the materials, participants, method, and data collection is referred to here as well.

### 3.4.1 Materials

The researcher created a website built specially for the replication and emailed the link to participants. The first webpage on the site presented a confidence test. This test had the form of a Likert scale test (see Appendix I), and it was introduced in the replication study considering one of the main objectives in the original study was to increase participants' confidence when confronting authentic materials. Johnson & Heffernan (2006) however, never really tested for confidence. Hence, I decided that in order to measure participants' confidence, they should be tested prior to and after the treatment to see whether there was any actual difference in confidence levels. After taking the pre-confidence test, subjects were automatically directed to the performance pre-test. After taking this pre-test, the online platform automatically directed subjects to the pages containing the readings and the trailers. In the last reading activity, and after watching the video, subjects were requested to follow the link at the end of the page in order to take the post-test and the post-confidence test. A link to the retention test was planned to be emailed to subjects seven days later, on the day of the test itself.

All trailers for the replication study were manually and individually selected taking into account age of participants, cultural appropriateness (for instance words or images related to certain topics such as profanity, violence or religion could not be portrayed in the videos) and audio clarity (see Appendix III for the list of trailers). Trailers were embedded in the study web pages. The portions of the trailers used in the tests were edited and embedded using YouTube.com. All trailers that were selected for the study had a slow-paced narration, and audio voice was clear and not particularly difficult to comprehend for non-advanced L2 Spanish learners. For consistency, all trailers were in Latin American Spanish since this is the variety their teachers spoke in class. On average, ten vocabulary items per trailer were selected as target words. Target vocabulary consisted of words that were considered to be difficult to comprehend without the teacher's assistance. Trailers were automatically shown only after the participant had read the reading section and had answered all ten questions correctly.

Figure 3-1 below shows a screenshot of a portion of a webpage with an example of five of the questions participants had to answer in order to watch the trailer. Figure 3-2 shows a screenshot of the trailer being played with subtitles. Notice that the subtitles are only shown here for illustration purposes, but they were blocked when participants were watching the videos. The subtitles in the screenshot are automatically created by the online video player, and they show some mistakes. The proper trailer text extract of that portion of the video without mistakes and with an English translation appears below the figure. A full screenshot of a webpage containing the reading portion, the questions, the answers section, and the trailer can be found in Appendix IV.

5- Iron man quiere----- al mundo.  
 proteger  atacar  venganza

6- Un país es una -----.  
 mundo  nación  vida

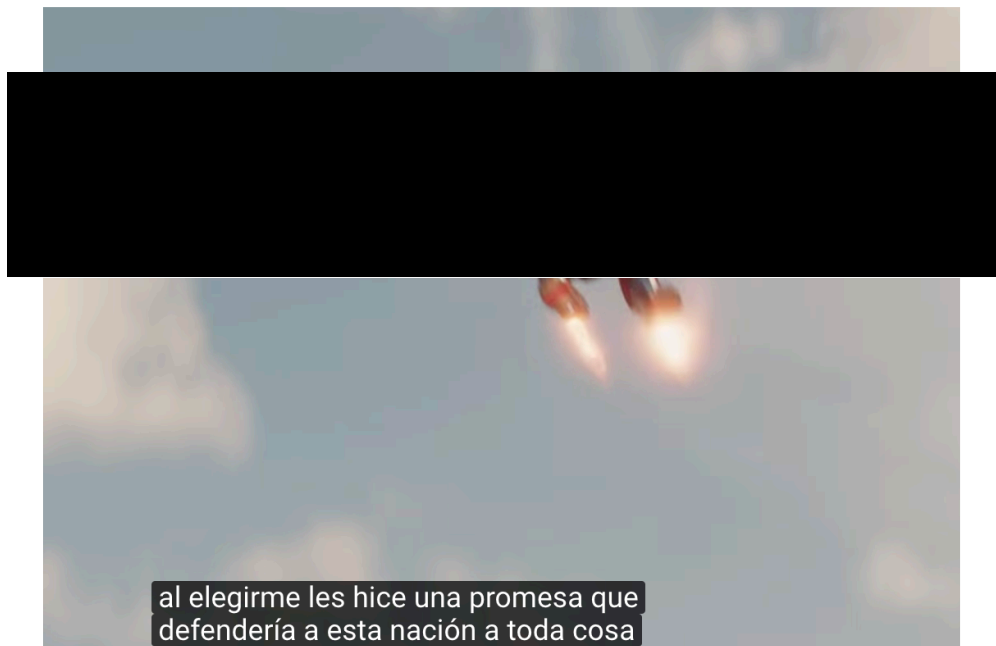
7- El Mandarín no quiere a Stark y quiere -----.  
 trabajo  venganza  miedo

8- Tony necesita -----para detener al Mandarín.  
 refuerzos  venganza  secretos

9- Tony tiene una ----- muy bonita llamada Peper.  
 nación  vida  novia

10- Los robots van a ----- al mundo de los ataques.  
 detener  llegar  defender

**Figure 3-1: Screenshot of five questions in one of the webpages of the replication study**



**Figure 3-2: Screenshot of a trailer being played with subtitles**

*al elegirme les hice una promesa, que*

*defendería a esta nación a toda costa*

‘when I was elected, I promised to

defend this nation at all costs’

Readings were purposely created for the project. Each reading was approximately 150 words in length and was related to the trailer that appeared below on the same page. Most of the vocabulary items used in the readings were taken from the corpus of RAE’s (Real Academia Española, 2014) 5000 most frequent words, and were appropriate for the participants’ level of Spanish. There were on average ten target vocabulary items per reading with a mouseover interaction that showed the definition in English. All target words appeared in bold type and underlined in the readings. When participants brought the mouse cursor on top of a target word, a pop-up window showed the English definition for that word. The meanings of the target

vocabulary items were reduced to the ones appropriate for the trailers. Figure 3-3 shows a portion of a trailer with the target words in blue, and with the mouseover function activated on the word *trabajo* ‘job’.

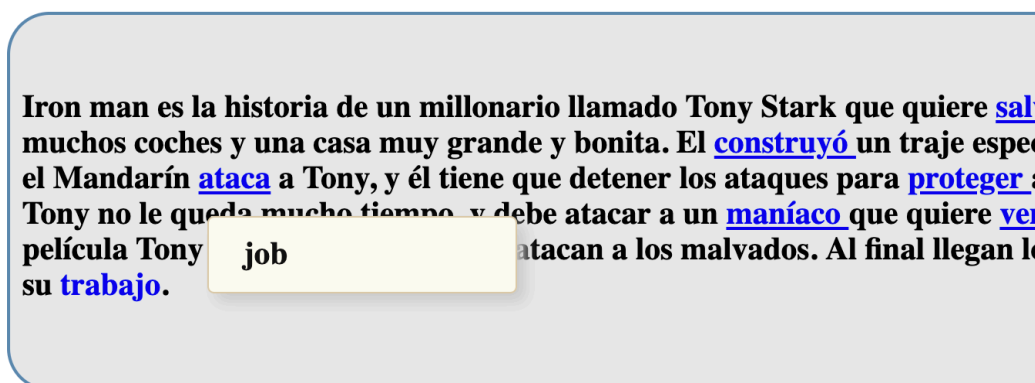


Figure 3-3: Screenshot of a reading with mouseover function activated

Similar to the original study, items appeared in different sections to ideally enhance learning and retention. In the replication, each target item appeared in at least one reading, one question section, and one trailer, with at least three repetitions in the whole project. The context surrounding the target vocabulary items served as contextual clues to help with comprehension.

### 3.4.2 Participants

Participants in this study consisted of a convenience sample of 50 L2 Spanish students. The subjects (aged 14 to 16) were in grades nine through twelve (in the American schooling system), of mixed gender and from seven different classes. Each class had an average of seven students and consisted of learners of similar Spanish abilities. Each class that participated in the study received instructions in English, remained in their usual classroom, in the presence of their usual teacher, and had the sessions during their regular Spanish class time. Each class had two Spanish lessons a week (Monday/Wednesday, or Tuesday/Thursday). Each participant worked on their own laptop computer and used headphones. Participants were not allowed to use

paper or online dictionaries or digital translators. All students in this study attended an international school in Qatar. There was a large number of different nationalities (23) in a relatively small number of subjects, participants had a wide range of mother tongues, and they all used English as their common language.

### **3.4.3 Method**

The replication began with a teachers' meeting to organize how the sessions would be delivered during the intervention. Since participants were in seven different classes, two teachers and the experimenter were in charge of leading the sessions. In the meeting, the experimenter provided a quick induction to the two teachers who learned about the objectives, organization of the project and what was expected from them. Teachers were told that they would meet their usual students at their regular class times, they only had to ensure participants were on task and followed the links, and they could not offer any help with comprehension.

Data collection in the replication was planned in a similar way to the original study, following a single group experiment with a pre/post-test design. To test acquisition of new vocabulary items, participants took a 20-question pre-test to measure previous knowledge of target words. The same test was provided immediately after the last learning session as a post-test to check student learning. A sample question found in the test is shown below in Figure 3-4 (for a screenshot of the full test see Appendix II).

1- Un ----- es una ciudad pequeña.  
 fantasma  mundo  pueblo

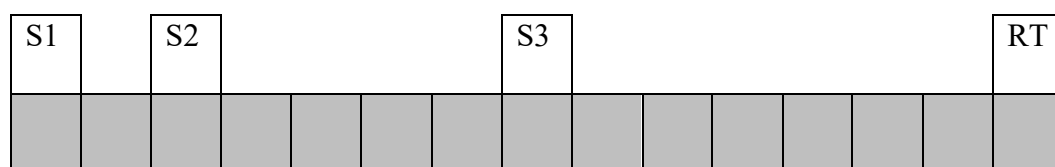
**Figure 3-4: Sample question of the replication pre-test**

This replication also included the confidence test described in the previous section, that was provided prior to and after the intervention. Data collection finished seven

days after the last learning session when participants took a retention test (which was the same test participants had taken as pre- and post-test) to investigate how much vocabulary was retained.

The overall intervention occurred in four sessions over a period of two weeks during participants' regular Spanish class. During the first session participants were explained that they would be part of a research project and that they could opt not to participate as their grades would not be affected. Next, the teacher in charge of the class read a set of instructions which consisted of a brief introduction to the project, duration, activities involved, and participants' roles (full set of instructions are shown in Appendix V). Later, participants proceeded to take the tests of the day (pre-confidence test and performance pre-test) followed by two readings with their corresponding set of ten questions and trailers each. During session two and the first half of session three subjects finished all of the readings and watched all of the trailers. During the second half of session three participants took the performance post-test and the confidence post-test. Seven days later, participants took the retention test.

The figure below shows the distribution of the learning sessions (S1, S2, S3), and finishes with the retention test (RT) on the last day of the study. Notice that the shaded area is divided in different cells, and each cell represents a day of the week. For example, between session one (S1) and session two (S2) there was one day in between. The retention test (RT) occurred on the seventh day counting from session three (S3).



**Figure 3-5: Replication study timeline**

### 3.5 Results

This section analyzes the results of each one of the three groups of tests individually (learning, confidence, and retention). This section will also refer to the special case of the retention test and the reasons for not including its results in the final discussion.

#### 3.5.1 Vocabulary tests

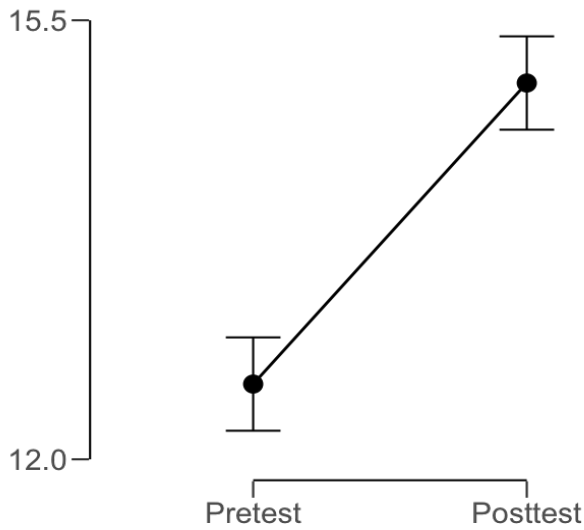
The vocabulary tests contrasted target vocabulary knowledge before and immediately after the intervention. To analyze the results, a paired *t*-test (since data was collected from the same group) was conducted. A first analysis of the data, as shown in Table 3-1 revealed that results from 50 participants (as represented by the N value) were collected in each test. The scores (as represented by their mean) of the post-test (M=15) were higher than those of the pre-test (M=12.6). SD showed that pre-test scores were more spread apart than post-test results (pre-test SD= 3.642, post-test SD= 3.037). This revealed that probably participants had different language levels prior to the study. A lower post-test SD showed that knowledge of target vocabulary among participants was more homogeneous after the intervention.

	<b>N</b>	<b>Mean</b>	<b>SD</b>
Pre-test	50	12.60	3.642
Post-test	50	15.00	3.037

**Table 3-1: Replication pre-test/post-test results**

Figure 3-6 below shows the difference in mean scores between vocabulary pre- and post-test presented on Table 3-1 above. The graph clearly shows an improvement in performance in the post-test mean. This probably implies that after the intervention participants had increased their knowledge of the 20 target words being tested.





**Figure 3-6: Mean difference between performance tests**

The next step in the *t*-test analysis was to check whether the samples were normally distributed, since if they were, a parametric test would be used, otherwise a non-parametric option would be preferred. A Shapiro-Wilk normality check revealed that samples were normally distributed at  $p=0.089$  (see Table 3-2). In this kind of test, a P value that is less than 0.05 is said to be statistically significant and therefore not normally distributed.

		<b>W</b>	<b>p</b>
Pre-test	- Post-test	0.960	0.089

**Table 3-2: Test of normality (Shapiro-Wilk)**

A parametric Student's *t*-test was run to contrast results. Findings below (see Table 3-3), show that there was a significant difference between test scores, and results of the post-test ( $M=15.00$ ,  $SD=3.037$ ) were higher and more homogeneous than results

of the pre-test ( $M=12.60$ ,  $SD=3.642$ ),  $t(49)=-9.165$ ,  $p<.001$ . Finally, a Cohen's  $d$  value of  $-1.296$  shows a large size effect.

		<b>t</b>	<b>df</b>	<b>p</b>	<b>Cohen's d</b>
Pre-test	- Post-test	-9.165	49	< .001	-1.296

**Table 3-3: Paired samples  $t$ -test**

This section revealed that there was a clear difference in scores between the pre-test and the post-test, and that this difference was highly statistically significant with a large effect size. The following section will compare results from the confidence pre-test and post-test in a similar fashion.

### **3.5.2 Confidence tests**

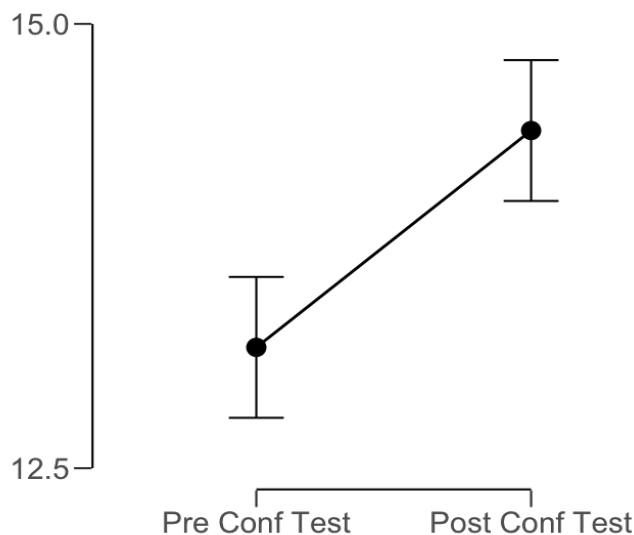
The confidence tests checked how confident participants felt when working with authentic materials. Subjects took a pre-confidence test prior to the intervention, and a post-confidence test at the end of the last learning session. In order to measure confidence levels in participants, the replication study included a pre- and a post-Likert scale test. To compare results between the two tests, another paired  $t$ -test was conducted.

A first analysis of the data, as shown on Table 3-4 revealed that the scores (as represented by their mean) of the confidence post-test ( $M=14.40$ ) were higher than those of the confidence pre-test ( $M=13.18$ ).  $SD$  showed that pre-test scores were more spread apart than post-test results (pre-test  $SD= 2.775$ , post-test  $SD= 2.695$ ). A lower post-test  $SD$  showed that there seemed to be a more homogeneous feeling of confidence towards authentic materials at the end of the intervention.

	N	Mean	SD
Pre-confidence test	50	13.18	2.775
Post-confidence test	50	14.40	2.695

**Table 3-4: Descriptive statistics of both confidence tests**

Figure 3-7 below shows a graphical representation of mean scores of pre- and post-confidence tests. The graph shows that participants were probably more confident dealing with authentic materials towards the end of the intervention.



**Figure 3-7: Mean difference between performance tests**

A Shapiro-Wilk normality check revealed that samples were normally distributed at  $p=0.147$  (see Table 3-5). A parametric Student's  $t$ -test was run to contrast results and findings (see Table 3-6) show that there was a significant difference between test scores, and results of the post-test ( $M=14.40$ ,  $SD=2.695$ ) were higher and more homogeneous than results of the pre-test ( $M=13.18$ ,  $SD=2.775$ ),  $t(49)=-4.374$ ,  $p<.001$ . Finally, a Cohen's  $d$  value of  $-0.619$  shows a medium size effect.

		W	p
Pre Conf Test	- Post Conf Test	0.965	0.147

**Table 3-5: Test of normality (Shapiro-Wilk) for confidence tests**

		t	df	p	Cohen's d
Pre Conf Test	- Post Conf Test	-4.374	49	< .001	-0.619

**Table 3-6: Paired samples *t*-test for confidence**

This section revealed that there was a clear difference in scores between the pre-test and the post-test confidence tests, and that this difference was highly statistically significant with a medium effect size. The following section will refer exclusively to the retention test which deserves special consideration since it did not happen as planned.

### 3.5.3 Retention test

This section refers exclusively to the retention test that was planned as part of the general intervention, but since it did not take place as planned, its data was not included as part of the general discussion in this chapter. A description of the test, its results, conclusions, and reasons for not including it in the main replication discussion are described below.

The test was a replication of the pre-test and the post-test and was taken by only some participants seven days after the last learning session. Since at the school where the project took place a few days before the test there was a dangerous situation that was beyond the control of the researcher, several of the participants were absent to school

on the day of the test. As a consequence, only 27 participants (out of the 50 that had taken the pre-test and the post-test) took the retention test.

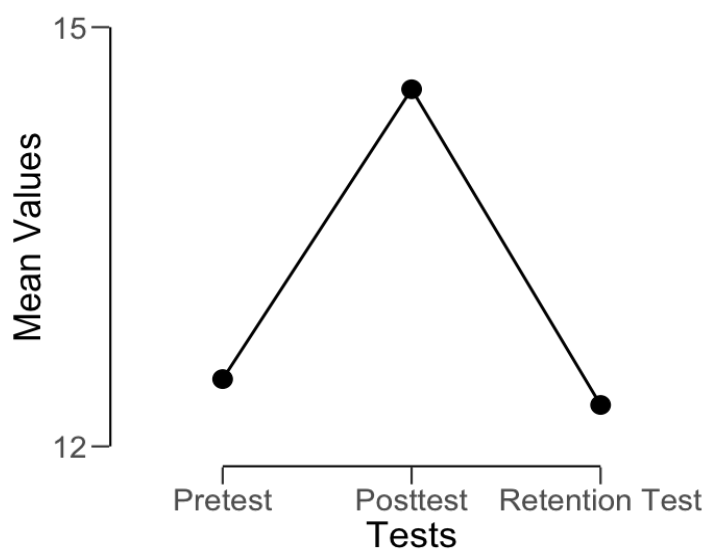
Considering that almost half of the subjects were absent to school, and since the ones that took the test showed very little interest to continue to work on the project, commitment and motivation was extremely low. At the same time, some results could be questionable since some participants seemed to randomly select answers to test questions. Therefore, the results shown on Table 3-7 were not included in the final Discussion section of this chapter. Results of the retention test are still presented here since they were part of the replication study, but its findings should be analyzed with caution.

Notice that data shown on Table 3-7 differs from data presented above since here scores included only the 27 participants that took all three tests. A one-way repeated measures ANOVA test was conducted to analyze results further.

The first comparison between test scores can be seen in Table 3-7 which reveals that the mean values of the post-test are higher than the other two ( $M=14.56$ ). The results of the retention test ( $M=12.30$ ) do not seem to differ much and are even lower than those of the pre-test ( $M=12.48$ ). Results of the post-test are the most homogeneous since they have the smallest standard deviation ( $SD=2.940$ ). This difference in results can also be appreciated in Figure 3-8 further below. In order to investigate whether this apparent difference between the tests was statistically significant (since the samples came from the same group) a repeated-measures ANOVA test was conducted. The test was used to calculate  $F$  which represents the variance between the samples. The repeated-measures ANOVA test requires that the equality of the variances between the samples should be met. This is called the sphericity condition which should be checked in advance. Therefore, a Mauchly's test of Sphericity was run. Table 3-8 shows results of such test which presents the value of  $P$  as 0.949. This means that there was no significant difference in the variances as  $p>0.05$  and therefore the ANOVA test could be run without applying any sphericity corrections to it.

	<b>N</b>	<b>Mean</b>	<b>SD</b>
Pre-test	27	12.48	3.227
Post-test	27	14.56	2.940
Retention test	27	12.30	3.232

**Table 3-7: Mean scores of the three vocabulary tests**



**Figure 3-8: Mean values of replication tests**

	<b>Mauchly's W</b>	<b>p</b>	<b>Greenhouse-Geisser <math>\epsilon</math></b>	<b>Huynh-Feldt <math>\epsilon</math></b>
Scores	0.996	0.949	0.996	1.000

**Table 3-8: Test of Sphericity**

The ANOVA test (see Table 3-9) shows that combined results of all tests were statistically significant  $f(2,52)=29.44$ ,  $p<.001$ . These results revealed that a significant difference existed between groups, but in order to know exactly where the differences were, a post-hoc Bonferroni test was also run. The post-hoc analysis (see Table 3-10) revealed that both the difference between pre-test ( $M=12.48$ ,  $SD=3.227$ ) and the post-test ( $M=14.56$ ,  $SD=2.040$ ), and the difference between post-test and the retention test ( $M=12.30$ ,  $SD=32.32$ ) were significant at  $p<0.001$ . The difference between the pre-test and the retention test was not significant as  $p=1.000$ . The same table also showed Cohen's  $d$  ( $d$ ) scores which revealed large effect sizes in results contrasting the pre-test and the post-test, and the post-test and the retention test. However, there seemed to be a small size effect when contrasting the pre-test and the retention test.

	<b>Sum of Squares</b>	<b>df</b>	<b>Mean Square</b>	<b>F</b>	<b>p</b>
Scores	84.96	2	42.481	29.44	< .001
Residual	75.04	52	1.443		

**Table 3-9: Replication ANOVA test**

		<b>Mean Difference</b>	<b>SE</b>	<b>t</b>	<b>Cohen's d</b>	<b>p<sub>bonf</sub></b>
Pre-test	Post-test	-2.074	0.337	-6.150	-1.184	< .001
	Retention	0.185	0.320	0.578	0.111	1.000
Post-test	Retention	2.259	0.323	6.997	1.347	< .001

**Table 3-10: Post-hoc comparison of replication tests**

The aim of the retention test was to check how much information was still kept seven days after the last learning session as a result of the intervention. This could have helped obtain a general idea of whether the methodology employed had indeed

promoted retention of target items. Unfortunately, due to the inconveniences mentioned above, conclusions cannot be made so straightforwardly. However, even when results above have too many threats to validity to be included in the final discussion (e.g., number of participants, lack of interest by the participants), there are still interesting findings worthwhile analyzing. It is still worth noticing that while there is a difference in learning between the pre-test and the post-test, the retention test showed the lowest mean scores. Even more interesting is the fact that retention scores were even lower than scores participants showed prior to the intervention in the pre-test. Even when participant cooperation was low during the retention test, this still shows that at least with the participants tested, target vocabulary was probably not being kept in memory as initially expected.

### **3.6 Discussion**

This section analyzes findings from the replication itself, and then it will continue to explain how those findings relate to the original study and to previous research. Later, since this replication was also used as a pilot study, the third section will refer to observations and conclusions made regarding the development of the intervention. Finally, this section finishes by providing guidelines for further work and suggestions.

#### **3.6.1 Analysis of results from the original vs. the replication study**

Considering this was an approximate replication study there are some differences between the replication and the original study that need to be addressed. However, even when the replication and the original study were very similar in some way, it is still interesting to see that despite the many differences between them, findings, in general, were still very similar.

To start with, similar to the original study, participants scored higher in the post-test at 0-day RI, revealing that vocabulary learning had indeed occurred. A smaller standard deviation in the replication post-test ( $SD= 3.037$ ) showed that target vocabulary knowledge was more homogeneous across participants, probably suggesting that at least some of the learning had occurred thanks to the intervention. There was a general mean score of 15 (in 20 questions with 50 participants) in the



post-test taken immediately after the last learning session, representing an average group score in the replication study of 75%.

This increase in group average results shown in the post-test was found in both studies. The group increase in the original study showed a 16% word gain in section one, with contextual based questions, and 8.1% word gain in section two with questions based on video clips averaging a 12.05% as a whole combining both sections. The replication study, on the other hand, showed a word gain of 12% in the vocabulary post-test combining contextual and video questions in the same test. This reveals that both studies had almost the exact same word gain. Even when similar findings are generally expected from a replication study, it is still surprising considering both studies differed in many crucial aspects. This is of particular relevance, then, considering the fact that participants in both studies had very different demographics and were learning two different languages. Participants in the original study were university Japanese students taking L2 English, while participants in the replication study were high-school students from multiple nationalities studying Spanish L2. Since English nowadays is a world lingua franca, it could be argued, that considering the English language presence in Japanese media (McKenzie, 2008) and the push from the government to learn English (Yashima, 2009), participants in the original study should in theory, have had greater exposure to the target language in comparison to those in the replication study (high-school students learning L2 Spanish in Qatar). This could potentially have led to larger vocabulary gains for original participants, making comprehension of authentic materials easier for them, in comparison to participants in the replication itself. Interestingly enough, that was not the case. It would also be interesting to see whether similar findings would still hold in another replication study where the learning method is modified, for instance by presenting and rehearsing target words through the use of digital flashcard platforms or when the authentic materials used are not movie trailers but rather short stories or magazine articles, for example. An increase in comprehension of magazine and newspaper articles would be of major importance for a course such as Spanish Ab Initio.

Another aspect that also resulted as similar across the two studies is poor learner motivation. This is of particular importance considering that the two studies differed widely in mainly two factors. The first of those is target language difference between the studies. Again, considering that university students in the original study were learning L2 English, it can be easily argued that those students would be intrinsically more motivated to learn a language that could be very useful to them in several aspects of their everyday life. In contrast, participants in the replication study were taking L2 Spanish, which is not seen as necessary as English nowadays, therefore motivation to learn it could be lower in contrast to participants in the original study. The second of those factors that could have influenced subject motivation is participant age. Several studies claim that, in general, university students tend to be more motivated than high-school students (e.g., Pintrich & Schunk, 2002; Tüysüz, Yildiran, & Demirci, 2010), since, in theory, students opt to enter university and choose their course of studies following their personal interests. High-school education, however, is generally seen as compulsory, and therefore, there is a feeling of rejection shared by most students for that same reason. This topic is referred to as intrinsic motivation by Pintrich & Schunk (2002), and it drives people to work on tasks because they find them enjoyable. In a study about university students and their motivation towards science (Tüysüz, Yildiran, & Demirci, 2010), the authors concluded that students have a positive attitude when they willingly chose the subject of studies. Therefore, what seems to call the attention is the fact that both studies showed rather low levels of motivation despite the difference in target language and student age. This lack of motivation was also mentioned by previously analyzed studies (e.g., Erbes et al., 2010; Goossens et al., 2012; and Gryzelius, 2016) which emphasizes the need to plan strategies accordingly to improve motivation and engagement, especially in a longitudinal field study like the one suggested in this thesis.

The final aspect in this section is the environment where the studies took place. The original study seemed to have taken place in a very controlled environment (the authors do not mention any external intricacies interfering with the development of the study). The replication study on the other hand, was a pure field study in a natural high-school environment where several issues had to be sorted for the study to take

place as planned (e.g., student absences, teacher collaboration, tardiness, motivation). It can be argued that participants in the replication study could have been more distracted with their environment which could have contributed to lower scores. However, and similar to the previous paragraph, these differences between the studies did not prevent findings from being similar across the original and the replication study.

To conclude, despite major differences between both studies, findings were actually replicated. Findings from the replication agree with the original authors' conclusion that the online activity helped boost learners' comprehension of authentic materials. At the same time, in both studies, vocabulary tests showed that participants had increased their knowledge of target words. The confidence test in the replication also indicated that participants felt more confident when dealing with authentic materials at the end of the intervention.

### **3.6.2 Results of the replication study against the literature**

Findings from the replication were further contrasted against different studies analyzed in the Literature Review and their word gain in post-tests taking place immediately after the intervention at 0-day retention interval (RI) (see Table 3-11 below). Notice that the table shows a 68.7% gain for Johnson & Heffernan (2006), which is the average of the combination of both post-tests in the study (reading: 72.5%, video: 64.9%). Erbes et al. (2010) shows no interstudy intervals (ISI) since there was only one learning session in the study, with a final exam immediately following the session. Finally, the replication study shows two to five ISI days considering there was no fixed interstudy interval between the learning sessions.

Article	Lab /field	Study sessions	ISI (days)	RI (days)	%
Bahrlick (1979)	L	3	1	0	87%
		6	1	0	98%
Bloom & Shuell (1981)	F	3	1	0	84.25%
Johnson & Heffernan (2006)	L	9	7	0	68.7%*
Fitzpatrick et al. (2008)	L	20	1	0	95.33%
Erbes et al. (2010)	F	1	-	0	36-52%*
Replication study	F	3	2-5	0	75%

**Table 3-11: Word gain comparison among different studies at 0 RI**

Table 3-11 above unveils interesting facts. To start with, Erbes et al. (2010) is the article showing the lowest word gain but it was also the one where the authors reported several issues regarding teaching and learning and poor participant rapport with the teachers leading the lesson. Apart from this article, Johnson & Heffernan (2006) and the replication study are the two lowest scoring studies on the table. Although there are several reasons that could contribute to low scores (e.g., poor methodology, low student engagement), it could be argued that these two studies, are, at the same time, the only ones with interstudy intervals longer than one day. Therefore, considering that they obtained the lowest scores at 0-day RI, this seems to agree with findings from Bahrlick (1979) stating that learning is faster at shorter lag times between learning sessions (ISI). This same notion is also shared by other scholars (e.g., Bahrlick et al., 1993; Bahrlick & Hall, 2005; Cepeda et al., 2008) suggesting that learning is higher at shorter interstudy intervals, but retention is shorter. In the same way, learning is apparently lower at the beginning at larger interstudy intervals, but information can be retained for longer periods of time. This second claim can unfortunately not be corroborated with the two studies discussed here, since Johnson & Heffernan (2006) did not test retention, and results from the replication retention test are not reliable. A final issue that could have also contributed to relative low word gain in the original and the replication study is the number of

exposures to target words. In both studies, for the most part, participants were exposed to each word three times. Irrespective of the number of learning sessions in the study the fact that some target words had a low number of rehearsals (three times) agrees with findings from Fitzpatrick et al. (2008) stating that more rehearsals contribute to better learning and retention.

Similar to the original study, this replication also aimed at increasing participants' confidence when dealing with authentic materials. This feature was not tested in the original study, but this replication, however, included a pre- and post-confidence test to obtain more precise information. Therefore, as seen in section 3.5.2 above, when contrasting results from both replication confidence tests, findings show that participants seemed to feel more confident when dealing with authentic materials by the end of the intervention.

To sum up, this replication study had three main initial objectives. On the one hand, it aimed at improving student learning which would eventually lead to better comprehension of authentic materials (movie trailers). It also aimed at improving retention of target vocabulary through repeated exposures. The third objective of this replication was to raise participants' levels of confidence towards authentic materials. Test results show that two of the main objectives were actually met. Participants seemed to increase their knowledge of target vocabulary and improve their understanding of authentic materials (although vocabulary learning scores seem to be rather low, especially when contrasting them to similar studies as seen in Table 3-11 above), and test results also show that participant confidence grew towards the end of the intervention. Due to the unexpected events affecting the retention test, the retention goal could not be met as test results were not reliable.

### **3.6.3 Replication as a pilot study**

This replication also served as a pilot study to test how a research project on spaced repetition (SR) could be carried out in a similar environment to where the main project would take place. While the intervention was in progress the researcher collected impressions and reflections through informal interviews and discussions with teachers and participants. The teachers leading the lessons provided their

impressions to the researcher and there was a final oral discussion after the last confidence test where students shared their reflections of the project as well. The topics below are based on researcher's observations, teachers' comments and participants' reflections.

To start with, the researcher, teachers, and participants all seemed to agree that spaced repetition (SR) seemed to work well in the study. Participants seemed to remember words (they had learned at the beginning of the project) when working on readings or trailers appearing towards the end of the intervention. There was a general sense of achievement in participants, and they seemed to enjoy the fact that after doing the readings they could understand words from the trailers. Participants also mentioned that in the trailers at the end of the study they were able to recognize words that had appeared in previous reading activities or in another trailer.

The researcher noticed that one of the most important topics that arose from the replication and appeared as crucial for educational field study success was school context. In the replication study it was confusing and frustrating at times to see that some teachers were not fully committed to the project, or that in some occasions participants would take longer periods of time to get ready to start the project. There were also a few other inconveniences that had to be dealt with on the spot to allow the project to run smoothly (e.g., low internet connectivity, computers failing to start, freezing or crashing, fire drills, students pulled out from class unexpectedly).

Finally, a recurrent issue that can also negatively affect the development of a field study is participant motivation. The first and the last period classes of the day proved to be less effective for motivation and subject engagement. Participants also seemed to really enjoy watching the trailers, but they did not appreciate the readings and tests in the same manner, therefore, their willingness to go through them rapidly diminished.

To sum up, although the intervention took place with only one major inconvenience (the one affecting the retention test as mentioned above), there were several other issues that still needed to be attended to in order to ensure the intended progression of the study. This suggests that in a field study several unexpected issues can take place

that could attempt against the planned progression of the project. Therefore, even when there are unexpected events that might still take place without warning, it might be useful to prepare contingency plans to allow for sudden accommodations without compromising the course of the intervention.

All in all, the pilot study served its purpose by providing data to be used as guidelines for future research. Other topics related to future research, in particular to how the original study can be expanded even further are referred to below.

#### **3.6.4 Further work and suggestions**

The replication study aimed at expanding the original study by Johnson & Heffernan by collecting some extra data (through confidence and retention tests). However, there are still some issues that can be addressed to complement the original study and further comprehend how spaced repetition (SR) can contribute to learning and retention, how learning can be improved in such a study, and most importantly how poor motivation can be avoided.

To start with, the original study might have benefited from a more thorough organization of the interstudy sessions and also with the inclusion of a retention test at the end of the treatment. According to scholars discussed before (e.g., Bahrick, 1979; Cepeda, et al., 2006; Cepeda, et al., 2008; Küpper-Tetzl, et al., 2014) retention of knowledge largely depends on the retention interval and interstudy session (RI/ISI) combination. Therefore, this study could be expanded by firstly deciding on how long after the last learning session information is planned to be retained (retention interval), and according to that, decide on the optimal lag that will separate each learning session (interstudy session) when information will be repeated. This however, is not so straight forward considering that there are still some discrepancies regarding the exact time to repeat given a certain retention interval. As discussed in the conclusion of the Literature Review, there are some scholars that state that the retention interval and the interstudy interval should be equal in length (e.g., Bahrick, 1979 and Lotfolahi & Salehi, 2017), while others claim that the interstudy interval should only be a portion of the retention interval (e.g., Cepeda, et al., 2008 and Küpper-Tetzl, et al., 2014). Therefore, expanding the original study by testing

retention at different intervals after the last learning session, could even contribute further and add more data regarding the optimal length of the interstudy interval given a certain retention interval as well.

Secondly, the original study and the replication resulted in word gain that seems to be rather low, especially in comparison to other similar studies (see Table 3-11 above). Learning could arguably be improved by teaching vocabulary explicitly through vocabulary lists or flashcards, since there is substantial research suggesting that this method can help acquire large amounts of vocabulary quickly (e.g., Thorndike, 1908; Bahrick & Phelps, 1987; Bahrick, et al., 1993; Fitzpatrick et al., 2008; Schmitt, 2008; Nakata, 2011, 2015). This could be later followed by the methodology employed in the original study with the readings and the videos. For example, the target words of the project could be presented to the participants as a printed bilingual list. Later participants could be directed to an online flashcard platform to create their own flashcards using the words from the list. Participants could later practice the target words engaging in some interactive activities or games offered by the online platform. After this, subjects could be asked to work on the readings, followed by the questions and watch the trailers at the end. Subsequent learning sessions could concentrate specially on activities focusing on retrieval from memory as it appears to enhance vocabulary learning and retention (Nakata, 2017). This could be achieved by asking participants to go over the flashcards at the beginning of the session trying to remember the meaning of each one and then check for confirmation.

Finally, in both studies (original and replication) student motivation was not as high as expected, therefore it seems sensible to think of different strategies to keep students' interest high for the duration of the project. It is interesting to see, however, that at least in the replication study, researcher's observations and participants reflections revealed that subjects were eager to watch the trailers and enjoyed the progression from poor to better comprehension of the words that appeared in them. It seems that using movie trailers for motivation was well received by participants. The fact that they had to read, answer questions and take tests was probably not.

Therefore, based on the ideas of two articles analyzed in the Literature Review (McLean et al., 2013; and Milliner, 2013) that participants had to sit a final exam at



the end of the intervention, seemed to keep subjects motivated when working on the research project. According to those two studies, participants were willing to learn the target words since they would help them obtain a better grade in an external exam after the intervention. As a consequence, at least in a field study, it seems that participants are more motivated to actively participate in the intervention if target words are part of the curriculum and participants' grades depend on their knowledge of them.

To conclude, Johnson & Heffernan (2006) and the replication study could easily be expanded to obtain more conclusive data regarding their main objectives. The use of flashcards to teach the target vocabulary before starting with the readings could improve learning, and the addition of tests after the learning sessions could provide important data regarding retention of target vocabulary. Finally, the inclusion of target vocabulary in tests affecting course grades, at least in field studies, could arguably increase participant engagement and motivation.

The next section provides the final conclusions of the replication study and presents motivations for the main project of this thesis.

### **3.7 Conclusion**

This section refers to findings in the replication study and how they motivate the main project proposed in this thesis. The replication provided evidence that recycling target vocabulary does improve learning. It also demonstrated that there are still some issues regarding learning, spaced repetition (SR), and motivation that need to be addressed in order to enhance retention of information, especially in longitudinal field studies like the one proposed here. This section then will refer mainly to those three main topics that arise from the replication and will need to be addressed in the main project of this thesis: Retention, Learning, and Motivation.

To begin with, although the replication failed to provide reliable data regarding a retention post-test, if a study aims at improving retention, then, a spaced repetition plan should be created beforehand. As it was mentioned above, just spacing vocabulary learning sessions alone cannot guarantee that information will be kept in memory longer in time. According to Bahrick (1979), if spacing between learning

sessions is too short, then vocabulary is forgotten quickly after learning. In contrast, if vocabulary is repeated at very long intervals between sessions, then it might be forgotten, and it might need to be re learned again. This refers back to one of the main concerns of the Literature Review: how long the gap between learning sessions should be, and what the optimal interstudy interval (ISI) should be for a given retention interval (RI). Considering that there is a lack of conclusive data in that matter, the main project of this thesis hopes to contribute to the field by providing data that could help understand the issue further.

The second topic that needs to be addressed is learning. Both studies (original and replication) showed very low word gain (see Table 3-11) in comparison to similar studies. This emphasizes the need for proper learning in order to guarantee that information is acquired properly and that it can be retained. This topic was discussed partially in the previous paragraph specially referring to lags between learning sessions. This paragraph, on the other hand, concentrates specially on the teaching methodology employed for better learning. Fitzpatrick et al. (2008) demonstrated that large amounts of vocabulary can be acquired quickly through the use of bilingual vocabulary lists. This idea can even be further expanded through the use of digital flashcard platforms that seem to be more attractive to participants than lists on paper, and could in turn, enhance learning, retention and engagement, even in field studies, as seen in McLean et al. (2013), Milliner (2013) and Gryzelius (2016).

Finally, both the original and the replication study revealed that poor participant motivation and engagement can be a major concern in a field study. This therefore highlights the need to conduct a study where participants are willing to participate so data obtained can be relied upon. The fact that participants in the main project will need to pass an external final exam (IB Spanish Ab Initio) seems to be a good motivation on its own, as it was demonstrated in McLean et al. (2013) and Milliner (2013) by having participants take external exams after the intervention.

To sum up, this replication concludes that further research is indeed needed regarding the use of spaced repetition (SR) to improve retention. At the same time, teaching methodologies ensuring that vocabulary is learned, memorized, and revisited

periodically at the right intervals arose as a major concern necessary for consequent long-term vocabulary retention. Finally, although increasing motivation is not one of the main goals of this thesis, it still appeared to be a crucial factor to bear in mind in a longitudinal field study. The main research project of this thesis will touch upon these issues. The following chapters in this thesis will introduce the methodology employed in the main project, results, discussions, and conclusions.

## Chapter 4: Methodology

### 4.1 Introduction

This chapter begins by providing a general overview of the main focus of this research project. The methodology employed to answer the research questions will be presented later, followed by a detailed recount of every single learning session and data collection session.

The replication study introduced in the previous chapter focused on spaced repetition (SR) as a method to improve vocabulary acquisition, to enhance vocabulary retention, and to help comprehend authentic materials. Findings from the replication study also emphasized the need for further research regarding spaced repetition since although participants showed improved learning, results were relatively low in comparison to previous similar studies. Another aspect that also emerged from the replication was the lack of motivation shown by participants. Even when the original (Johnson & Heffernan, 2006) and the replication study had employed movie trailers, in part to increase subject motivation, there were instances when participants were not interested in working on the project. Finally, the replication also served as a pilot study that helped with the planning and organization of this research project to enhance learning and retention, and to improve participant motivation.

The relative low retention scores of Johnson & Heffernan (2006) and the replication study could be attributed to the lack of a systematic agenda of interstudy and retention intervals. This topic arose as one of the most important issues in the Literature Review since for spaced repetition (SR) to enhance vocabulary retention, the retention interval (RI) and the interstudy intervals (ISI) should be carefully determined before the actual teaching begins (e.g., Bahrick, 1979; Cepeda et al., 2006; Cepeda et al., 2008). This RI/ISI optimal combination, however, is still to be established since there are discrepancies among the same scholars in relation to the ideal combination of both. In the same chapter, it was also discussed that the length of time information is retained is important, but it is crucial also to determine how much of that information is actually retained as well. The final issue that arose from the Literature Review was that more field studies with high-school students are needed (most previous spaced

repetition studies had university students as participants) in order to further understand how vocabulary retention works in young learners.

## **4.2 Overview of the Study**

This thesis was conceived with the notion that in the school where the research project took place, IB Spanish Ab Initio students tend to forget previously learned vocabulary by the time they sit their course final exam. This is particularly important in courses that extend over two years as is the case of the Spanish Ab Initio course. In order to enhance vocabulary retention, this research project aimed at strengthening vocabulary acquisition by testing participants at different intervals on their target vocabulary knowledge. The overall project design consisted of a target vocabulary preselection test, a pre-test, eleven learning sessions, plus a post-test, and two delayed post-tests.

## **4.3 Methodology**

### **4.3.1 Materials**

This subsection describes the procedure for selecting target vocabulary for the project and conditions and processes to include them in the study. Together with the target items, this subsection also focuses on teaching resources employed and reasons for selecting them.

#### **4.3.1.1 Target vocabulary**

This subsection will explain the reason for selecting 100 words for this project. This decision was based on a combination of a pre-assumed average word gain for an Ab Initio course (see below), and previous research findings (Fitzpatrick et al., 2008).

In the school where this study was carried out, on average there were 38 school weeks yearly. The Ab Initio course took three lessons of one hour each week, making it a general 114-contact-hour course per year. Taking into account the multiple and different school eventualities that prevent a lesson from being taught as planned from time to time (e.g., unannounced fire drills or events, unexpected visitors, field trips), 100 contact hours was considered as the annual average for this type of course. Based on their general L2 language teaching experience, and on their knowledge of the Ab

Initio course, curriculum designers at the school had estimated a rough annual word gain of 600 words, at a rate of six words per contact hour. I then estimated that for a study focusing mostly on vocabulary acquisition a higher word gain could be expected. Therefore, it seemed that, with the appropriate method, a total of 100 items over eleven learning sessions could be successfully learned and retained in this type of course. This resulted on an average word gain of just above nine words per contact hour. The nine (plus) word gain per contact hour proposed for this project fitted perfectly between the traditional six word gain for an Ab Initio course, and the fifteen word-gain per lesson as discussed in Fitzpatrick et al. (2008). Fitzpatrick et al. (2008) had concluded that in the study with only one subject, a highly motivated university professor had managed to learn a long list of target words (300) in 20 days, at a rate of 15 new words a day. This high word gain cannot be easily expected in a longitudinal field study with young learners. Therefore, considering all of the differences between Fitzpatrick et al. (2008) and this project, the researcher decided that nine words per day over eleven sessions could be challenging enough and, yet attainable, for these participants in the present study.

A final note on the challenge that target words would pose to participants is that the researcher also purposely chose nine words to be learned per learning session trying to avoid a ceiling effect. A ceiling effect takes place when most of the examinees obtain the maximum possible score in a test (Ary et al., 2018). This means that if the task is not challenging enough, most of the subjects would obtain high scores, probably limiting participants in their learning, not showing how much they could have actually learned. Also, a ceiling effect would cause test results to be skewed and not be normally distributed.

#### **4.3.1.2 Final 100 target words**

The selection of the 100 target words for the project followed two steps that will be described in detail below. First, the researcher manually selected 120 words from an IB dedicated list, and then participants took a test to define the final 100 target words actually used in the study.

The researcher first selected 120 words (see Appendix VI for a complete list of all 120 words with test results) from the IB Spanish Ab Initio suggested list (IB, 2002). The list presents a number of topics to be taught in the course together with suggested vocabulary that could be covered to succeed in course assignments and exams. The list is not mandatory, nor exhaustive, but merely a guide for teachers to be able to plan their course. An example of a topic (*el individuo*) with suggested words found in the list is:

*el individuo* ('the individual')

*información personal* ('personal information')

*datos personales* ('personal data'): *nombre* ('name')

*edad* ('age')

*sexo* ('sex')

*estado civil* ('marital status')

The pre-selection itself was based on the researcher's experience teaching the course, and on materials purposely created to teach the course by the teachers at the school. All materials to introduce new vocabulary during every unit were based on the IB list. Typically, at the beginning of the unit the students received lists of words and phrases to be learned during the unit. The teacher would provide the students with a Spanish-only printed copy (for a small sample of how new words were presented to students in a unit see (Appendix VII) and would go over it in class. The teacher would ask the students to work in pairs and translate the words into English using a dictionary. The teacher later would read the words and phrases to students (together with the English translation) so they could check for errors. The lesson generally continued with students making sentences and writing dialogues using the Spanish words provided. For the research project, the researcher selected words that did not normally appear in

the materials used to teach the units in order to avoid words that were generally taught explicitly in class. This way the researcher made sure that none of the words pre-selected as target words would be taught explicitly in a regular course lesson (these lessons were regular course lessons that were not related to the research project itself). However, some words included in the target list of the research project were vocabulary items that would generally be taught explicitly in class. When the time came to work with those items in class during regular course lessons the researcher skipped those items to avoid interfering with the development of the project. An example of this was the word *película* ('movie') that would normally be explicitly taught in class in a word list towards the end of the Ab Initio course with the topic of Free Time. Therefore, when the time came to teach the word *película* explicitly in a regular course lesson, the researcher purposely deleted that word from the list given to participants in order to avoid giving extra attention and exposure to that word outside of the project.

Although in regular course lessons target vocabulary was normally first introduced to students in its basic form (verbs were not conjugated, and nouns and adjectives were kept in their neutral gender and number form, for instance, *alumno* 'student', *curar* 'to cure'), base conjugation and gender and number inflections were also taught to students so they could use the words in context. The exact same strategy was applied in the project, and apart from appearing in the flashcards in their base form, nouns and adjectives also appeared in their inflected forms in different project activities. Since participants had already been exposed to Spanish noun and adjective gender and number agreement and to basic verb conjugations (for more detail on this see 4.3.2 below) this rarely caused any trouble during the project, but if there were questions on these topics, the researcher provided some help.

In order to make the final selection of 100 items for the study, a test which consisted of an online multiple-choice activity was purposely created on Quizlet (<https://quizlet.com>). The researcher created an online set of 120 bilingual Spanish-English flashcards. The system then automatically provided four English distractors (taken from the Spanish-English flashcards) per headword. Figure 4-1 below shows a



sample of three questions taken from the actual test in which for each Spanish word four English distractors were provided.

### 120 Multiple choice questions

1. alegre

- always
- water
- happy
- horse

2. asignatura

- spring
- subject
- studious
- secondary

3. cabello

- hair
- small
- horse
- later

### Figure 4-1: Sample of the 120-word pre-selection test

Participants in the experimental group took the multiple-choice test to check their knowledge of the 120 initial words. Only the experimental group was tested on this since school administration had especially requested the researcher to avoid distracting teachers and students from other courses as much as possible. Also, participants from the control group would be at the same level and would know in general the same words.

Test results were used to collect the final target-word list. The 20 words with the highest correct answer scores were discarded and the remaining items composed the

final 100-target word list (see Appendix VI for the 120-word list and their test scores, and Appendix VIII for the final list of 100 target items for the project).

Once the target items had been decided, the following issue to deal with was resources to be used for testing and learning. Those resources are introduced next.

#### **4.3.1.3 Teaching resources**

The introduction and teaching of target vocabulary took place during the learning sessions, and mostly through the use of bilingual (Spanish-English) digital flashcards. Although there were some paper-based activities, the majority of the exercises participants engaged in were technology based and different online platforms (described below) were used at different times.

Bilingual flashcards were employed for teaching since findings from several scholars suggest that flashcards can help acquire large amounts of vocabulary quickly and efficiently (e.g., Thorndike, 1908; Bahrick & Phelps, 1987; Bahrick, et al., 1993; Fitzpatrick et al., 2008; Schmitt, 2008; Nakata, 2011, 2015). At the same time, the use of digital flashcard platforms seems to enhance learning and motivation (e.g., Milliner, 2013; McLean et al., 2013; Gryzelius, 2016), as they could be fun to work with, and seem to have a multi-sensory appeal (Stutz, 1992). Finally, learning and retention seem to be enhanced also when learners create their own flashcards (instead of working on a ready-made set) since encoding appears to be stronger by doing, than by just looking and repeating (Slamecka & Graf, 1978, Nakata, 2017). In line with this, Kosslyn & Smith (2006) states that there seems to be a major difference in the strength of encoding when information is transformed into memory representations. For instance, if at the time of encoding we pay attention and are actively involved, the process of encoding information will most likely be strong. Finally, forgetting may also appear as a consequence of poor encoding for instance in the case of lack of attention when an event is happening. Recent findings seem to support the claim that active engagement also favors learning as in Webb & Piasecki (2018) learners of English as a foreign language seemed to improve their vocabulary acquisition by writing words down.

Just as in the test for selecting the final target words for the study, in order to work with the online digital flashcards Quizlet was also used in the project. Quizlet allows users to create their own sets. The system also automatically creates several interactive activities (e.g., matching, filling-in-the-blank) and games using the bilingual sets. The online platform offers learners different opportunities for learning and rehearsal, as well as, the opportunity to learn in a more amusing manner by playing the games. Figure 4-2 below shows a sample of a flashcard on Quizlet made by one of the participants. The flashcard shows the word in Spanish. The English translation can be seen if the user clicks on the flashcard itself. To the left of the flashcard the system offers different interactive activities that can be played using the bilingual sets.



**Figure 4-2: Quizlet flashcard sample**

Participants in this project were already very familiar using Quizlet in their everyday life (using it at home and in school in different subjects), which helped with the development of the sessions since subjects did not have to spend time familiarizing themselves with the system. In order to create their flashcards, participants received a printed copy of the 100 target words (Appendix VIII) and then the researcher checked

for errors in the sets (for full description of the way participants created their flashcards on Quizlet see 4.4.2 below).

The second online platform used in the study was Quia (<https://www.quia.com/>).

Although Quizlet was used for the preselection test of the target words, all other tests were delivered through Quia, since this platform provides better data collection and statistical tools to analyze results. While learners generally use Quizlet for self-study, since the platform is more user friendly and graphics make it catchier, educators have more options for teaching if using Quia. This online platform is mainly geared towards educators where they can create their own activities for their courses, specially through a paid subscription. Although the interface seems rather dated since its design is too basic, Quia offers plenty of data to analyze test results in depth.

Finally, while students access the system and work on the different activities, the system also records time-on-task and results. Figure 4-3 and Figure 4-4 below show a sample of the test in learning session four, and a sample of how a score distribution analysis for that test was presented to the user.

## Session 4 - 50 words test

1. solicitar (1 point)

- to walk
- to request
- foreigner

2. periodista (1 point)

- bag
- magazine
- journalist

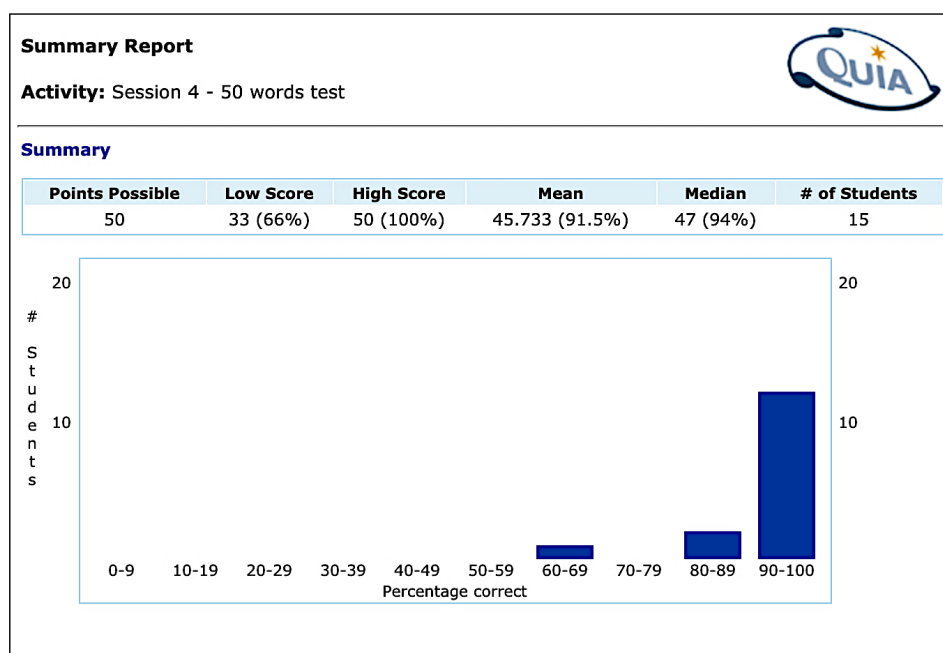
3. parientes (1 point)

- broken
- relatives
- food

4. panadería (1 point)

- week
- bakery
- nurse

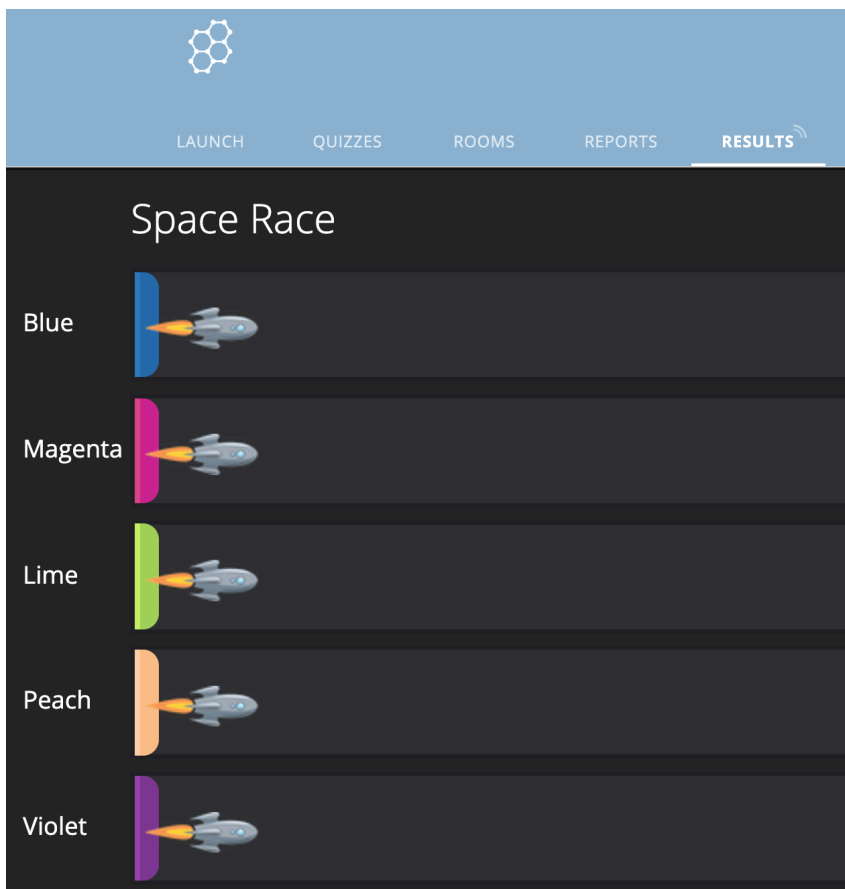
Figure 4-3: Quia test sample



**Figure 4-4: Quia statistical sample**

Finally, a third platform called Socrative (<https://socrative.com/>) was also used in two sessions (see 4.4.9 and 4.4.12 below) to play an interactive game (Space Race). Socrative was used in this case because participants had already played the game in other subjects and they really enjoyed the friendly competition. Socrative is an online response system that offers immediate feedback. The platform seems to be very effective for formative assessment and student understanding can be tracked in real time. The Space Race feature itself is an engaging interactive game. To play the game the educator first creates the questions and answers in the system. Later when the teacher selects the Space Race option, the system takes the questions and answers created before and prepares the game through a multiple-choice activity. Students can access the game on their portable devices by going to Socrative.com and then entering a code leading to the game. Each student is assigned a rocket with a different color for the game (this feature is of particular relevance since the student's name is left anonymous which could avoid potential embarrassment and bullying in class). The game begins when the system shows questions on the player's individual screen, and every time the player answers to a question correctly their rocket moves forward. The

first rocket to reach the finish line wins the game. Although each player can only see the questions and options on their individual screen, the system also allows for another view of the race where the teacher can see the rockets moving along the screen as students answer the questions. Learners seem to really enjoy playing the game as they see their rockets moving along and racing against their peers, especially when the teacher's screen is shown to the whole class using an overhead projector, for example. Figure 4-5 shows an example of a Space Race about to begin.



**Figure 4-5: Screenshot of Space Race on Socrative**

Finally, apart from the different online platforms mentioned above, participants also worked on paper. For example, in order to help with memorization, in certain activities such as when revising the target words or in reading comprehensions

participants were encouraged to take notes and write the words on paper as it seems to enhance vocabulary learning (Webb & Piasecki, 2018). At the same time, some of the activities assigned by the researcher were also based on paper in order to simulate the format of the readings to be found in the final exam. 4.4.4 below shows an example of a reading activity in which participants had to read a story and work on it.

### **4.3.2 Participants:**

As stated in 4.1 above, most of spaced repetition (SR) studies have used adults as participants, which shows that there is a need to test spaced repetition with young learners. The majority of those studies have been conducted in laboratories, and those that took place in an actual classroom are short in duration. This project instead, used high-school seniors as participants during regular school hours.

To start with, since the project took place during regular class time, subjects were at the same time students taking a course required for graduation and participants in a research project. Careful thought was put into the organization and delivery of learning sessions and testing, since participants could not be deviated from their course requirements considering their grades were crucial for university admissions.

All participants in the study, aged 16 to 19, attended a private international school in Doha, Qatar. The large majority of the students in Spanish Ab Initio class were expats, which always led to a course with about fifteen students per class with over ten different nationalities and several different mother tongues.

By the time the project started, participants were close to finishing the first year of their two-year program. All participants began the course without any previous formal Spanish education, since in order to be admitted to the course they had to comply with the requirement of having no previous formal education in the target language (Spanish). Participants had three lessons of one hour in duration per week. The two-year course as usual started in September and finished nineteen months later in April the following calendar year. At the school where the project took place, during their senior year, students usually had a month of study leave, and sat the final reading and writing exams in May.



Considering the number of students interested in taking the course was close to thirty every year, by school rules, two classes were created of about fifteen students each. Each class would have a different teacher, but both cohorts would have their lessons on the same day and at the same time and would strictly follow the exact same curriculum. Therefore, by the time all participants took the project pre-test, they had all covered basic Spanish vocabulary on basic topics: physical descriptions of people and animals, colors, numbers, the house, the neighborhood and the school. They had also covered some basic Spanish grammar: present tense, and gender and number agreement of nouns and adjectives. Although planned to be taught later in the course, other topics in the curriculum also included food, health, sports, culture, celebrations, travel, environment, technology, and free time, as well as some other verb tenses: past and future tenses, and occasionally subjunctive. Due to the fact that some target words in the project required past tense in order to appear in context (e.g., *ayer*: ‘yesterday’), participants were introduced to Spanish simple past tense before they started this study. For instance, a week before the experimental group took the test with the 120 vocabulary items, a whole lesson was devoted to introducing participants in the experimental group to Spanish simple past tense. The lesson started with a group revision of Spanish present tense (participants had already learned the present tense about six months before during regular course lessons) and then the researcher explained orally to the class how the simple past tense of regular verbs and some irregular verbs was formed. Participants received three printed lists that consisted of regular verbs (those whose conjugation follow a certain fixed pattern such as *trabajar* ‘to work’) and also some irregular verbs that are important for learners but do not follow the same pattern as the regular verbs such as *hacer* ‘to do’. After receiving a printed copy of the lists (see Appendix IX), learners were asked to work in pairs and with the help of a dictionary use the verbs on the lists to write five sentences in simple present and turn them into simple past. After checking for errors orally, the researcher asked students to continue to work in pairs and write a whole paragraph (about 60 words) in simple past using a dictionary if needed. The lesson finished with learners handing in their writing to the researcher for correction.

For the project, subjects were divided into three main groups. There was an experimental group (EG) that underwent a spaced repetition (SR) treatment, while a control (CG) and a historical group (HG) were used for comparison.

Both the experimental and the control group were intact cohorts with 15 and 12 participants respectively who had been taking Spanish Ab Initio for about six months by the time this project started. Subjects in the control group were in the exact same situation as participants in the experimental group (i.e. they were students in their last two years of their secondary education taking Spanish Ab Initio, and their demographics was similar to that of the experimental group's). One of the main differences between the experimental and the control group was that participants in the control group were taught by a different teacher who followed the 'traditional' method of teaching the course (this is further explained in 4.3.3 below). The historical group, on the other hand, had 32 participants (composed of all of the students taking Spanish Ab Initio, that by the time this project started were in the second year of the course) who were about two months away from graduation and were one year ahead (academically) of the experimental and the control group.

The concept of the historical group was based on studies that used older data to contrast current results (Loudon, 2008), and control groups that received no instruction but still took tests to use as baseline (Mayer & Anderson, 1992). The purpose of the historical group was to obtain data from a cohort that was about to graduate by the time this project started. That data would be used as reference to have a baseline for performance in the pre-test as an average level of knowledge of the project's target words typical Spanish Ab Initio students would have at the end of the two-year course. The historical group was tested only once, before the project's learning sessions began, and on the same day the other two groups took their pre-test (see Appendix X for a complete timeline of the project including testing and learning sessions). In short, this data collected from the historical group would be later used twice to compare where the experimental and the control groups were at in relation to the historical group before the learning sessions began. The same data (coming from the historical group) would be used again later at the end of the learning sessions. This second time, the historical group's data would be contrasted against the

experimental and control group's post-test results (at 30-day RI) in order to see if there was any difference in learning between the groups.

Both, the control group and the historical group, were similar in that they took tests (only once in the case of the historical group) but did not receive any treatment. The main difference between those two groups, however, was that the historical group was a year ahead in the Spanish Ab Initio course, and that the control group took two tests, one prior and one after the learning sessions. The main purpose of having a historical group was to expand project's results to Spanish Ab Initio in general (at least at the school where the project was conducted), and not just contrast project's results to the mere difference between the experimental and the control group alone. For instance, if there were any differences or similitudes in performance between the experimental and the control group, this (difference) could be considered as pertaining to those two groups in particular. On the other hand, by having a baseline, results could also be contrasted against a wider and more general database that could be representative of historical data for Spanish Ab Initio in general at the school, and results could be expanded to a much larger group (rather than just the control group alone).

Finally, it is important to mention that since this research project involved researcher's own students, and the project took place outside of the UK, the appropriate ethics procedure was followed. Complete information regarding compliance with the ethics procedure can be found in Appendix XI.

Prior to the beginning of the project, the researcher met with school administration and provided details regarding the topic and importance of the study. At the same time, the researcher also explained that neither students, parents, nor staff members would be forced into participating and that all documentation regarding participants and school details would be anonymized. Finally, since the project would take place at the educational institution and during regular school hours, school norms and regulations would be complied with, together with local customs and traditions.

After this the researcher also met with two other Spanish Ab Initio teachers and explained the project to them as well. Teachers were told that their collaboration in

the project was optional and that they could pull out at any time if they felt uncomfortable or were no longer willing to participate.

Finally, on the first day of the project, participants received a note (Appendix XI) providing details about the project. They were also informed that they would be part of a research project (but their participation was optional and they could pull out at any time if they felt uncomfortable) and that the topics covered might help them obtain a higher course grade. However, participants were not informed of what groups they were in, what activities were part of the project, and they also ignored the duration of the study. Subjects were asked to bring the note home, discuss it with their guardians, and if anybody had any concerns they could meet with the researcher to clear any doubts.

The following subsection will describe the strategies followed in order to deliver the lessons, which included deciding retention intervals, the length of the interstudy intervals, and project schedule.

#### **4.3.3 Method:**

In line with the notion that the retention interval should be decided first to enhance long-term retention, this subsection goes from the general to the particular by firstly presenting the overall structure of the project before moving to the details of how each lesson was planned. Therefore, this subsection begins with the general organization of the study, and then it continues with the retention interval (RI) and interstudy interval (ISI) lengths preferred, together with the reasons for doing so. Finally, detailed information regarding tests, teaching methodologies and strategies are also provided below.

##### **4.3.3.1 General organization**

This subsection provides a general overview of the research project, and how it was generally scaffolded in order to merge it into an ongoing course as seamlessly as possible. The experimental group (the only group undergoing spaced repetition strategies) followed the school curriculum normally having three Ab Initio lessons a week. One lesson per month, however, would be devoted to the research project when

the corresponding learning session or assigned testing would take place. This lesson that was devoted to the research project tended to be the very first lesson of every month (in some cases, however, for external reasons the session had to be moved to a different day as detailed in 4.4 below) and participants in the experimental group worked on project specific activities. This unavoidably required a slight deviation from their course curriculum. For example, instead of working on the preestablished monthly topic of the Spanish Ab Initio course, the lesson focused on the research project where the online flashcards were used to rehearse the study's target words. This deviation, however, was actually partial considering the target words of the study would still be useful to pass the course requirements. At the same time, participants were still learning content related to their course curriculum and, whenever possible, similar activities to those of the course were also used in the project (e.g., reading comprehension activities for the project usually followed text types commonly covered in the Ab Initio course). Also (since the research project had a special set of target words to work with), in order to avoid interference between vocabulary learned during normal course lessons and the research project's target vocabulary, the researcher avoided explicit reference to any of the target words in any lesson that was not part of this study.

There were three instructors involved in this research project. The researcher was in charge of the experimental group, which meant that he led all sessions that required the participation of this group. The second instructor was the Spanish teacher of the class that acted as the control group in the project. This meant that the second teacher acted as proctor (monitoring participants) when the control group was tested, and he was also asked to teach the Ab Initio course to the control group as he usually would. That meant that he followed the usual (Ab Initio) methodology without modifying any of the content or teaching strategies. Both the researcher and the collaborator teacher were experienced educators that had been working together in synchronization teaching the Ab Initio course for several years. This helped with the planning of the study since both instructors were used to work synchronizing topics, activities, tests, and vocabulary to be taught in the course. Finally, in order to avoid any unwanted changes in the teaching methodology, details of how this intervention was implemented were not shared with the second instructor. The third instructor,

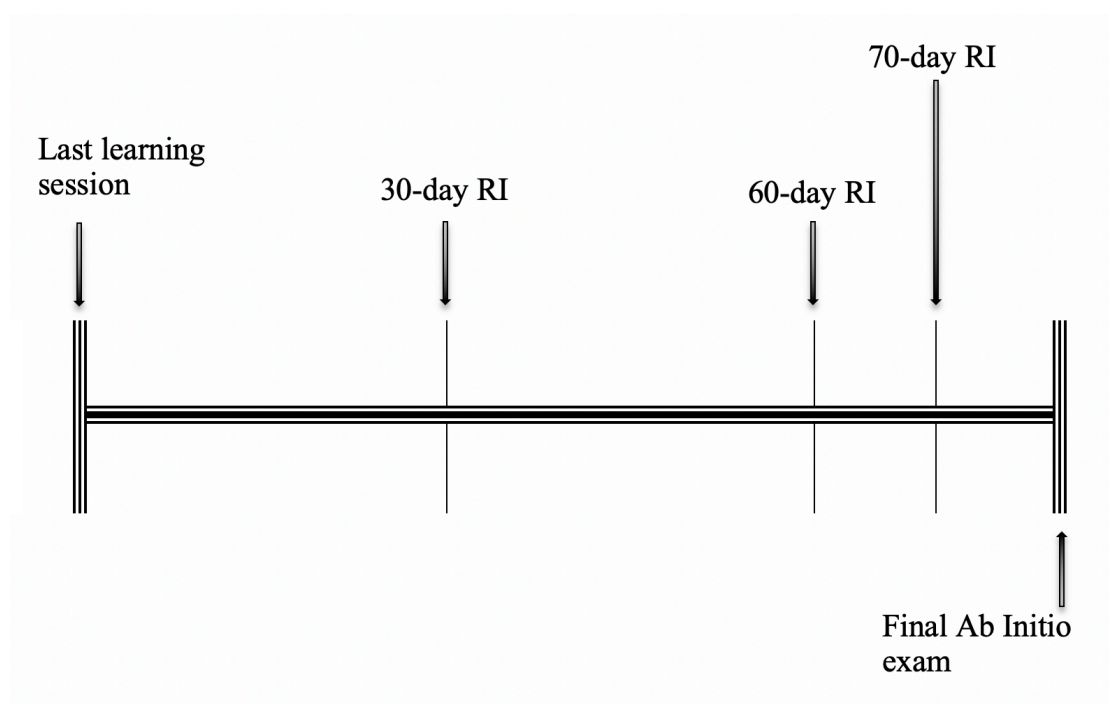
who was also another Spanish Ab Initio teacher at the school, was not involved in any teaching related to the project. The third instructor intervened only once by acting as a proctor of the historical group when they were tested for the study.

#### **4.3.3.2 RI/ISI lags, combination and study schedule**

This subsection describes the retention intervals' (RI) length, the optimal combination of the retention interval and the interstudy interval (RI/ISI). The overall study schedule including tests and learning sessions are also detailed here.

The length of the retention interval was based on findings from the Literature Review revealing that the field study with the longest retention interval in Table 2-13 was Gryzelius (2016) with a 39-day RI. This suggested that more field research was needed to comprehend how retention of vocabulary could actually take place in a longitudinal study with retention intervals extending over 39 days. In response to this, the researcher decided that a general 70-day retention period could provide enough data to analyze vocabulary retention as a result of the intervention. The 70 days were decided considering a suitable retention interval for the study, that could be accommodated seamlessly into the ongoing Ab Initio course, and to have the study finishing as close as possible to the Ab Initio final exam. Apart from those 70 days, the researcher decided to leave ten days between the final test of the project and the day of the course final exam. This was based on findings from the replication study where school context prevented the project to take place 100% as planned. Therefore, those ten days would be used, if needed, to accommodate any missing tasks due to unexpected events preventing the study to be conducted as originally planned. Thus, the researcher counted 80 days backwards starting from the day when the final Spanish Ab Initio exam would take place and established three post-tests to check vocabulary retention at different retention intervals. Those retention intervals were set based on the notion that they would provide enough data to analyze retention gains at different times after rehearsals had stopped. As a consequence, the researcher planned to test participants 30 days after the last learning session, 30 days after the first post-test, and finally 10 days later, with just ten days to spare before the Ab Initio final exam took place (see Figure 4-6). The 30-day RI lag was especially planned as it would help contrast project results against most of the studies listed on Table 2-13

above in which the retention interval extended from 30 to 39 days (e.g., Bahrlick, 1979; Sobel et al., 2011; Goossens et al., 2012; Küpper-Tetzel et al., 2014; Gryzelius, 2016; Lotfolahi & Salehi, 2017). The 60-day RI decision was simply based on the question of what the retention gain would be in twice the 30-day lag. Finally, the 70-day RI seemed to be the largest possible interval for this study (considering the days left in the school calendar to conduct the project).



**Figure 4-6: Project RI schedule**

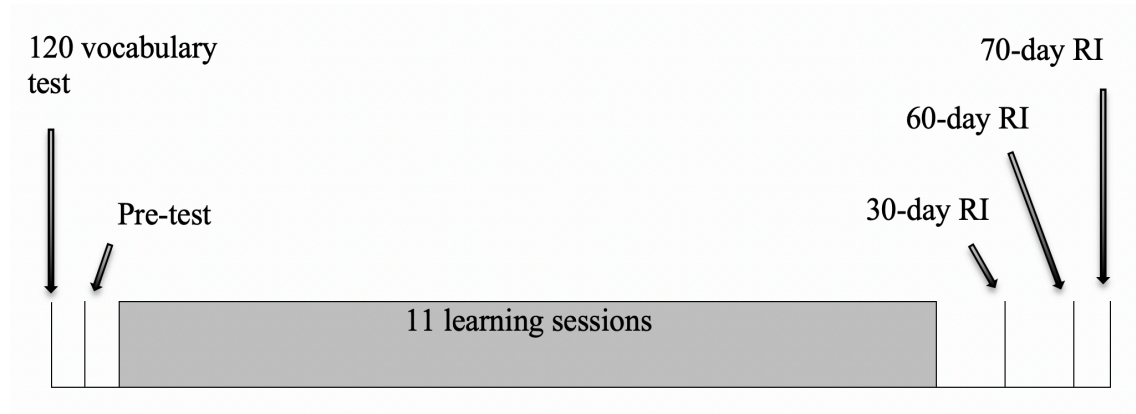
The next decision to be made was establishing the length of the interstudy interval (ISI). Two main options arose from the Literature Review regarding the most appropriate RI/ISI combination: the interstudy interval and the retention interval should be equal in length, or the interstudy interval should be only a portion of the retention interval (RI). In this study the first of the two options seemed to be more appropriate for three reasons. In the first place, an ISI that is equal in length to the RI can result in high retention gains, as seen for instance in Bahrlick (1979) which

reported a 95% vocabulary retention gain. Second, for a 35-day RI, repeating every seven or eight days (as suggested by Cepeda, et al., 2008) would not be very practical in a longitudinal field study with young learners. This would translate as repeating the same vocabulary once a week. This could therefore negatively affect the flow of the curriculum, and learners might probably lose motivation quickly. In contrast, at a given retention interval of about 30 days, repeating every 30 days would arguably be more practical in a natural teaching environment. In this case, a monthly vocabulary review would (ideally) not interfere with the flow of the curriculum. Instead, it could help both, enhance long-term vocabulary retention and keep the classroom environment as ecological as possible (which was actually the aim of this project). Repeating once a month would resemble what students would do in their usual Spanish Ab Initio language course. The third reason for equal RI/ISI length, is that when contrasting learning results (at 0-day RI) between repeating vocabulary at shorter and longer interstudy intervals, the difference is not so significant. For instance, Bahrick (1979) showed that, results from repeating every 30 days for six times were not dramatically lower than learning after six learning sessions at 1-day ISI. This revealed that even when learning at longer interstudy intervals seems to be lower, in comparison to learning at shorter interstudy intervals, there is not a particularly substantial difference between the two. Finally, a fixed 30-day ISI could also help provide further data to comprehend the RI/ISI relationship, by checking vocabulary retention at three different lags (30-, 60-, and 70-day RI). If after the treatment post-test scores are higher in the 30-day RI than in the other two retention intervals, this would show that the retention interval and the interstudy interval should be equal in length. If, in contrast, post-test scores are higher at 60- or 70-day RI, then results could probably support the claim that ISI should be a portion of the RI.

After deciding the most appropriate RI/ISI combination for the project, the final study schedule was planned. The complete study (see Figure 4-7) consisted of a pre-selection vocabulary test (120 words), a vocabulary pre-test, eleven learning sessions, and three post-tests. Notice that the tests mentioned in the figure below are the ones that directly referred to long-term vocabulary retention. There were three other tests in the project (the ones in learning session four, six, and ten). Those three tests took



place during the learning period and they were mostly conceived to monitor participant learning rather than to collect data about retention.



**Figure 4-7: Study organization, RI test days counted from the last learning session**

#### **4.3.3.3 Data collection**

This research project was conceived with two main goals in mind: to enhance long-term vocabulary retention while also aiming at the highest possible retention gain. Although robust vocabulary acquisition cannot be equated with long-term vocabulary retention (Nakata, 2017), without learning, vocabulary cannot be retained. In this sense, the strategies employed in this research project aimed at best possible learning. Considering the process of learning involves information being learned but some of it is forgotten between rehearsals (Bahrick, 1979), the tests given to participants during the treatment collected data to monitor their progress. For instance, tests in learning sessions four, six, and ten (see 4.4 below) were planned to assess how much learning had been taking place up to those points in the study. These tests were crucial for deciding also the progress of future learning sessions. If the researcher noticed that in general participants were showing poor acquisition, upcoming sessions would aim at reinforcing learning, for instance by focusing on encoding by rehearsals (for example through rote repetition) that could help with memorization of target vocabulary. Otherwise, sessions could continue as originally planned, for instance with target

vocabulary exposure through incidental learning (for example through reading activities or games). The rest of the tests, on the other hand, were used to collect data that would directly aim at answering the research questions. An example of those tests were pre- and post-tests that served to contrast participants' vocabulary knowledge prior to and after the treatment.

Eight different tests overall were administered in the study (see Table 4-1 for a list of tests, their purpose, and groups taking them). The first test was the 120-vocabulary test (see 4.3.1.2 for full details of the test) that the experimental group (EG) took prior to the treatment in order to determine the final 100 words to be used as target words in the study. The following test in the research project was the pre-test (see Appendix XII). The pre-test was the only one taken by all three groups in the study. This test consisted of 30 multiple choice questions, with a maximum possible score of 30 points. Each question had three options and two distractors (that were also part of the 100-word target word list). The questions were also formulated using target words. This means that knowledge of all target words was required in order to answer questions correctly. All questions and options were shown in a random order in every test, which meant that participants would not be answering questions in the same order and option words were presented differently in every question. A screenshot of a test question can be found below:

1. La /// adolescente es muy divertida y colorida. (1 point)
  - video juego
  - moda
  - folleto

**Figure 4-8: Screenshot of a pre-test sample question**

The following administered test was the one in learning session four (see 4.4.5). This test (see Appendix XIII) was only taken by the experimental group and the main purpose of it was to reinforce acquisition of the target words that participants in the experimental group seemed to be struggling with. The next two tests were the one in learning session six (which was the exact same one used as the pre-test) and the one in session ten (which was the same test administered in session four) that were used to monitor participant learning while the treatment was in progress. Finally, there was a post-test 30 days after the last learning session, and two delayed tests, 60 and 70 days, respectively, after the last learning session. All these three tests had the exact same format as the pre-test.

<b>Test</b>	<b>Purpose</b>	<b>Group</b>
120 vocab. test	To define the final 100 target words for the study	EG
Pre-test	To check target words knowledge prior to the treatment	EG, CG, HG
Session four	To test how much learning had been taking place, and to reinforce learning	EG
Session six	To test how much learning had been taking place	EG
Session ten	To test how much learning had been taking place	EG
30-day RI test	To check target words knowledge after the treatment	EG, CG
60-day RI test	To check target words knowledge after the treatment	EG
70-day RI test	To check target words knowledge after the treatment	EG

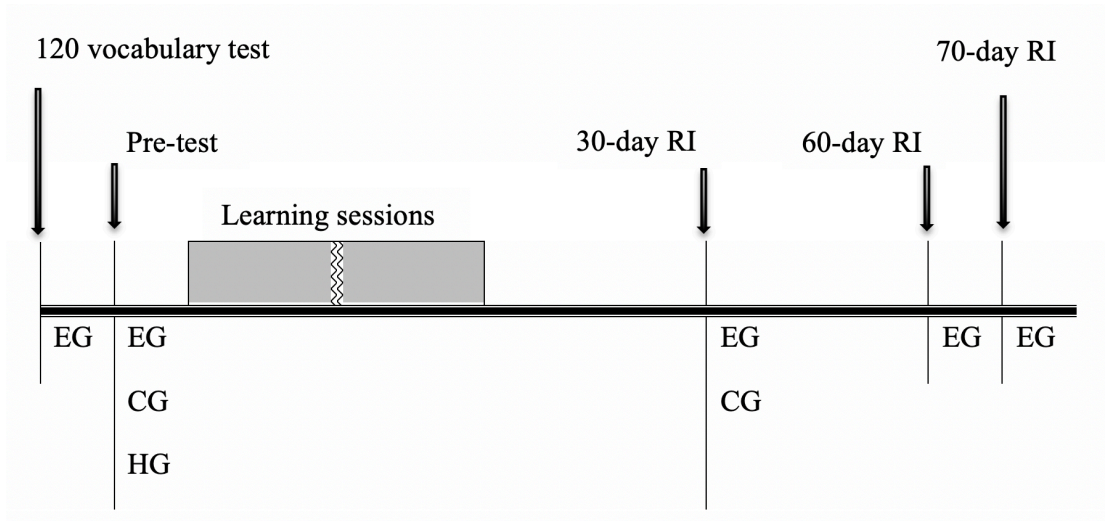
**Table 4-1: Tests in the study and their purpose**

Tests were administered to the groups according to their involvement in the project (see Table 4-2). The experimental group (EP) took all of the tests while the control group (CG) took only two tests (pre-test and post-test). Although the control group could have also taken the two delayed-post-tests (to reveal how much they were able to remember at 60- and 70-day RI), the researcher decided not to distract them from their usual course of studies.

Study Groups	Tests					
	120- vocab. test	Pre-test	Post-test	30-day RI	60-day RI	70-day RI
EG	√	√	√	√	√	√
CG		√	√			
HG		√				

**Table 4-2: Division of Participants by group and length of involvement in the study**

The historical group (HG), on the other hand, had very minimal involvement in the study by taking only one test (on the same day the other two groups took the pre-test. For a detailed timeline of the project by month see Appendix X). The reason for the historical group to take the test on the same day as the other two groups served two purposes. First, it provided a clear starting point of where each group was at an exact point in time prior to the beginning of the learning sessions. Second, since the 30-day RI test (that the experimental and control group took) happened almost exactly a year after the test taken by the historical group, it served to compare how far the experimental and the control group had reached in comparison to the historical group's baseline. This strategy aimed at answering research question one (Will spaced repetition produce different retention levels in current students in comparison to students who were taught using traditional teaching methods and graduated a year earlier?). Figure 4-9 shows a reduced timeline with a more graphical representation of each group and the tests they took. Notice that the shaded area in the table represents the learning sessions but this section has been reduced to show the tests more easily.



**Figure 4-9: Project timeline with test and group involvement**

So far, this Methodology chapter has focused on the scaffolding and the overall organization and delivery of tests and learning session. The actual teaching strategies employed in the study are described below.

#### **4.3.3.4 Teaching and learning**

Although a good RI/ISI strategy plus efficient teaching strategies could strengthen retention, there are factors that could negatively affect learning (and consequent retention) such as participants' lack of motivation (e.g., Johnson & Heffernan, 2006; Erbes et al., 2010; Goossens et al., 2012; Gryzelius, 2016). Subjects also showed low levels of motivation in the replication study discussed in the previous chapter. Thus, the researcher decided that two different strategies should be adopted aiming at enhanced motivation and improved learning. The first of those strategies was the use (as target words) of vocabulary that could help participants be better prepared for their Ab Initio course requirements, as well as the use of words in context which was also useful to pass course requirements. The second strategy was to include different activities in the learning sessions to keep subjects entertained, focused, and actively engaged. The first of the two strategies was based on findings from Milliner (2013) where the inclusion of a formal final assessment (TOEIC) seemed to keep subjects motivated. The second of the strategies was based on findings from Goossens et al.

(2012) that demonstrated that a variety in the learning tasks seemed to improve motivation that in turn could have also improved learning and retention. This also agrees with the notion that a variety of activities enhances encoding (Glenberg, 1979).

One of the main focuses of this thesis, regarding vocabulary acquisition, has been quick and effective learning. Thus, the majority of the learning sessions in the study included some explicit learning through rote repetition of all 100 target words through the use of digital flashcards.

Apart from studying target words explicitly in the study, participants also rehearsed words implicitly (for instance, through interactive activities and games using the digital flashcards) since this seems to reinforce learning and retention (Bahrick, 1979; Schmitt, 2008). Some learning sessions in the project also included practice with target words in context as these skills were required to answer questions in the test of study. This was also a way of connecting the research project to the Ab Initio course to keep participants motivated.

Finally, since the pre-test was administered five times, certain considerations had to be made to avoid the washback effect. Washback is defined as the influence that tests cause on educators. Apparently, teachers actually adjust their instructional habits since they want their students to do well in tests. It seems that learners' success reflects educators' own aptitude for teaching (Fournier-Kowaleski, 2005). Although the washback effect seems to be really strong in educational institutions where students' performance reflects the quality of the school, in this research project, washback was intentionally avoided aiming at more reliable research findings. For instance, results from all tests were never discussed during learning sessions. Second, although participants practiced with words in context during the learning sessions, none of the sentences used resembled those in tests. Finally, every time the researcher explained the meaning of a word in class, the examples were never similar to sentences found in tests.

The paragraph above showed the last portion of the Methodology chapter regarding general organization and planning of the project. The following section introduces all of the testing and learning sessions and how each one of them was actually delivered.

#### **4.4 Learning and testing sessions**

This section begins by presenting a summary of how each testing and learning session in which the experimental group was involved occurred and the topics the control group covered while the experimental group was going through the learning sessions. All testing and learning sessions of the study are introduced afterwards.

The first two sessions in the study were actually testing sessions in which the 100 target words were defined and when participants took the pre-test. All eleven learning sessions appear afterwards. The majority of the learning sessions involved explicit learning of target vocabulary followed by different interactive activities. At the same time, there were also several activities assigned where participants were exposed to target vocabulary implicitly, i.e. in an activity where target vocabulary was present but there was not explicit reference to it. These activities included games and readings, for example. After the learning sessions, participants met with the researcher three more times to take the assigned posttests 30, 60, and 70 days after the last learning session in order to check target vocabulary retention.

While the experimental group was going through the treatment, the control group, on the other hand, would follow the usual *Ab Initio* course curriculum as they normally would. This means that by the time the experimental group started with learning session one, the control group was covering the topic of City life. The complete list of topics for the duration of the study was the following: Country life, Health, Food, Holidays, Global issues, The Environment, Weather, Technology, and The Media. The teacher of the control group would typically focus explicitly on about 75 target words per topic, introduced mostly explicitly (see Appendix VII for an example of how target words were introduced in a unit) and through reading activities with a final revision at the end of the unit before a unit exam. A typical lesson would mostly consist of reading and writing activities, some videos, and the use of Quizlet as a digital flashcard platform. The lessons would also include some grammar topics such as simple present and simple past tense, as well as gender and number agreement of nouns, adjectives, and articles. The experimental group would follow exactly the same curriculum, work on the same activities and at the same pace, except for one lesson a month when they would go through the lesson devoted to the project. On the

same day, the control group would have a ‘light’ lesson to compensate for missed work (e.g., homework) or work on a special activity directed to improve participants individual weaknesses. For example, some subjects would do some writing, while others would work on speaking or reading activities. Participants in the experimental group would have the same opportunity to work on those activities once a week when they would be requested to visit the researcher during an hour in their class schedule dedicated specially for in-school tutoring or support.

#### **4.4.1 120-vocabulary test and pre-test**

The very first day of the study participants in the experimental group were asked to take the initial test that would define the 100 target words to be used in the research project. Two weeks later, the three groups (experimental, control, and historical) took the pre-test (see Appendix XII) that served as base data for this project. Participants took the test during regular class time, in their usual classroom, and in company of their usual instructor, which means that the groups were not together, and they took the test on the same day, and at the same time. The session started with the instructor explaining participants that they would be part of a research project, they received a study consent copy (Appendix XI) and were especially asked to let the instructor know if they were not willing to participate. Finally, the instructor continued to explain the format and rules of the test. Students could use the full lesson to finish the test, but if they finished ahead of time, they had to remain in the classroom and were allowed to work on any other activity.

#### **4.4.2 Session one**

This was the very first learning session and took place two days after the pre-test. Its main goal was to introduce participants in the experimental group to the target words so they could learn and have enough practice with them to ensure proper encoding in memory. In order to provide time for better learning and practice, session one was split into two different parts taking place on two separate days with one day in between.

Both parts belonging to session one started with participants receiving a (Spanish - English) printed copy of the 100 target words (Appendix VIII) and were asked to log



in to their Quizlet account and create their own digital bilingual flashcards. Subjects were asked to take only the first 50 word-pairs on the list in lesson one and continue with the remaining 50 the following lesson and create the remaining digital flashcards. The reason for using only 50-word pairs at a time was to provide enough time to create the flashcards, check for mistakes, and to have some time for revision by going through some of the interactive activities on the platform, such as the games, for example. The lesson started by asking participants to read through the printed list for ten minutes and try to memorize all 50 word-pairs. Later, subjects were asked to create online flashcards trying to remember the meaning of the words without checking the printed list. Participants were still allowed to check the list in case they were in doubt with any translation. The idea behind asking participants to retrieve information from memory rather than merely copying the list was based on the concept that retrieving information from memory enhances encoding and retention (Slamecka & Graf, 1978; Smith & Kosslyn, 2007; Nakata, 2017; Medina, 2019).

After creating their digital flashcards, and once the instructor had checked the flashcards had no errors, subjects were asked to continue to work on the words for the remainder of lesson one. Subjects were free to study the words by reading them aloud to themselves, writing them down on paper or by working on interactive activities, tests, or games that Quizlet had automatically created. However, participants were strongly encouraged to write the words down on a piece of paper to help them memorize the items since writing seems to improve retention (Webb & Piasecki, 2018).

The second part occurred the following calendar day and the remainder 50-word pairs were used to create the online flashcards. The activities were the exact same ones as in part one of this learning session. The only difference between the two parts was that at the end of the second part participants were requested to work on a printed list containing all 100 vocabulary items and to tick those items they still did not know the meaning of. The researcher collected the lists from participants at the end of the session and used the results to prepare the test for learning session four (see 4.4.5).

One participant was absent during the first lesson but compensated for it the next calendar day staying after school for fifty-five minutes. All participants were present in the second lesson.

#### **4.4.3 Session two**

Session two took place 31 days after the second part of session one and all participants were present. This session focused primarily on memorization of target words. Participants were asked to find their online flashcards and they were particularly requested to go over all 100 target words, one word at a time, trying to remember the translations before seeing them (before clicking on a digital button to flip the flashcard). Later participants were asked to study the target words for ten minutes in any way they felt most efficient.

For the remainder of the lesson participants were directed to three online interactive activities purposely created on Quia and they were asked to spend about five minutes on each one of them. The first task consisted of a matching activity where twenty random words at a time from the target vocabulary list were displayed on a table. Their shuffled translations into English were also displayed and the objective was to click on the appropriate translation (see Figure 4-10). The second activity was a game where participants answered questions and gained credits. Questions became more difficult as the game progressed, and the amount of credits won increased as well (see Figure 4-11). The final activity was the most challenging as participants were presented with a word in English and they had to type the Spanish equivalent. This proved to be the most difficult task of the session since most participants either made several spelling mistakes or did not know what the appropriate answer was (see Figure 4-12). Participants enjoyed the first two activities, and did very well overall, but many struggled with the third one. The final activity of the session consisted on the same activity participants did at the end of learning session one in which they had to tick those items they still did not know the meaning of.

aburrido	boring
adolescente	teenager
alegre	
almuerzo	
alquilar	
alumno	
asado	
ayer	
ayudar	
calle	
cabello	
bolsa	
billete	
biblioteca	
camarero	
cambiar	
carne	
casado	
cena	
cubiertos	

Figure 4-10: Screenshot of the activity to answer in English

curar

A to cut      B to drive

C to cure      D to eat

\$ 1,000,000  
 \$ 500,000  
 \$ 250,000  
 \$ 128,000  
 \$ 64,000  
 \$ 32,000  
 \$ 16,000  
 \$ 8,000  
 \$ 4,000  
 \$ 2,000  
 \$ 1,000  
 \$ 500  
 \$ 400  
 \$ 300  
 \$ 200

Hint Hint Hint

Figure 4-11: Screenshot of a game to practice the target words

1. old
2. video games
3. journey
4. to travel
5. glass
6. job
7. roof
8. card

**Figure 4-12: Screenshot of the activity to answer in Spanish**

Before participants were dismissed, for five minutes they reflected orally on their learning. Researcher observations and participants' comments showed that although participants seemed to remember most of the words, they still missed many, which showed that some learning had indeed taken place, but they needed further learning to encode concepts properly.

#### **4.4.4 Session three**

Session three offered opportunities for explicit and implicit learning, together with a group activity aiming at enhancing learning and improving motivation. This session took place 31 days after session two with all participants being present.

The session was divided in two parts. Just as in session two, this session also started with participants working on their online flashcards going over all 100 target words, one Spanish word at a time, retrieving English translations from memory. The second part consisted of a reading activity with a text using all 100 vocabulary items (see Appendix XIV). Observations and findings from session two showed that participants

still had some difficulties remembering the meaning of some target words. Therefore, vocabulary items either previously learned by the students in their regular course classes, or cognates were also included in the text to assist with comprehension. The reason behind this strategy was double folded. To start with, vocabulary familiar to participants was included in order to have subjects focus their attention on target words alone, and not get distracted with other unfamiliar items. The second reason was to have familiar words as surrounding context, near a target word, that could help decipher the meaning of it in case it was still unknown.

The reading activity was provided to participants as a means to reinforce exposure and repetition of target words implicitly, without the need to study them by memory. This activity was divided in two parts and it was created purposely for the project considering it presented a new and different exercise that could engage participants and keep them motivated. The task consisted on reading the text first individually for comprehension, and then work in groups to create *memes* (see below) to retell the story.

Subjects were explained that this activity was useful to them for two main reasons. To start with it provided practice to understand how target words behaved in context (e.g., verbs being conjugated, and nouns and adjectives agreeing in gender and number) as it was needed for the project. At the same time, the text was similar in style and format to the ones commonly found across the Ab Initio course, and it was also useful practice for reading strategies needed to succeed in the course.

During the reading section participants were given ten minutes to read the text individually unlimited times to comprehend as much as they could. They were allowed to use bilingual and monolingual dictionaries if needed. For the remainder of the lesson participants were split into four groups in order to create *memes* to retell the story through images.

*Memes* were used in this session since they provided opportunity for exposure to target words in a different and more entertaining way for participants. Per definition *memes* are amusing images with catchy texts that are easily shared and spread quickly (Dawkins, 2006; Brodie, 2009). Also, it seems that learners enjoy working with them

(Purnama, 2017) therefore, *memes* appeared to be an appropriate resource that could contribute greatly to learning, and student motivation and engagement. A sample *meme* portraying *estoy aburrido* ('I'm bored') created by participants during this lesson is shown below.



**Figure 4-13: Sample *meme* created by participants**

Before the session finished all groups presented their *memes* on the classroom screen. Just as in the two previous sessions, participants were asked to work on a list with all 100 target words ticking the ones they did not know.

#### **4.4.5 Session four**

This session took place 31 days after the previous one and consisted primarily of a test to check how much learning had taken place up to this point in the project. Starting in session one, at the end of every session participants had received a paper with a list of all 100 target words and were asked to tick those words they still did not know the meaning of. All words scored one point every time a participant ticked them. For this session, the 50 words with the most points, after the first three sessions, were used as target words for the test (see Appendix XIII).

The purpose of the test in this session was to determine how much participants had learned and how much they could remember after three learning sessions. This was the case considering one of the main goals of the project was to ensure maximal learning that could in turn lead to long-term retention and high retention gain of target words. At the same time, test results would provide useful information to adjust (if needed) content and delivery of future-learning sessions of the research project. For instance, if test scores showed that learning was not taking place as expected, then future sessions would concentrate more on memorization rather than mere implicit exposure.

The test was an online interactive activity created especially for this study on Quia. If participants made a mistake, they were provided with correct responses; therefore, apart from testing current learning levels, the test also helped participants with learning and memorization of target vocabulary at the same time.

The test consisted of a multiple-choice activity with 50 target words in Spanish. The target words employed were the ones subjects had selected as still unknown to them from learning session one to three. This time, distractors were in English, instead of Spanish, and they were all items guessed correctly by participants in all three previous sessions. Distractors were purposely selected in order to help participants guess target words correctly as much as possible. In case they selected a wrong response, the correct one was automatically provided.

One participant was absent, but he took the test the following day after school hours. All participants worked independently on their computers and were not allowed to use any help (dictionaries or group work was not permitted). All participants finished the test within 25 minutes and all of them were fully committed to it, with a positive attitude. Subjects seemed very confident on the task, as well. Once they all had finished the test, they were asked to refer back to the online flashcards they had created in session one, and played interactive games offered by the website.

Considering that tests results (see 5.6.1 below) and retention levels, up to this stage in the study, seemed to be very high and since learning was indeed taking place, learning

sessions continued to be as planned. This meant that participants seemed to be acquiring target words successfully which could eventually enhance retention.

#### **4.4.6 Session five**

Session five took place only nine days after the previous session. This was a special session with a much smaller interstudy interval due to an about 90-day summer break beginning ten days after the session. This session was particularly challenging considering participants would be out of school for an extended period of time, so the risk of forgetting was very high. Therefore, the main objective of this session was to reinforce memorization and proper encoding, to ideally avoid as much forgetting as possible.

The lesson started as always having participants working with the online flashcards individually retrieving the meaning of target words from memory. Later, subjects were asked to work in pairs (there was a group of three students). Each participant tested their partner on their knowledge of all 100 target words. The activity consisted on one participant reading the target words in Spanish one at a time. Their partner had to provide the English translation, taking note of words not translated properly, scoring one point every time they made a mistake. When finished, partners were requested to switch roles. Participants were told they would repeat the same activity at the end of the lesson, and they would be given a chocolate bar as a reward for improvement.

The next exercise consisted of participants working individually on Quizlet for ten minutes on the interactive activities using the flashcards going over as many activities as they could. For the remainder of the lesson, participants were asked to meet their partners again and repeat the group activity. At the end of the activity participants went back over the missed words, if any, and corrected their mistakes by finding the appropriate translation. As promised, participants, who had improved having a lower score in the previous activity, were given a chocolate bar as a reward by the researcher.

There was 100% attendance in this session, participants were fully committed to the tasks and the rewards proved to be very effective to provide extra motivation.



#### **4.4.7 Session six**

Session six occurred exactly 90 days from session five during the first Spanish lesson after the summer break. All participants were present and they took an online test to check retention and forgetting levels after their summer holidays. Results from the test in this session were used to monitor participant learning after five learning sessions and after a long summer break, and to adjust future learning sessions if necessary.

At the beginning of the lesson subjects were asked to follow a link on their email and take the online test on Quia. The test was the same as the pre-test. Just as in their pre-test, participants were not provided any feedback on questions they had missed or answered correctly as they were going to take the same test three more times at the end of the study. After completing the test, participants were asked to work individually on their online flashcards on Quizlet and, as before, they were asked to retrieve the meaning of the target words from memory before looking at both sides of the flashcards. For the last five minutes of the lesson, participants were asked to play any interactive activities they wished on the same website using the flashcards.

Results from the tests in this session (see 5.6.2) exposed the fact that most probably participants had forgotten some of the target words during the summer break. As a consequence, then, the following session concentrated especially on acquisition aiming at increased learning and improve retention levels.

#### **4.4.8 Session seven**

Session six provided useful data to determine research progress, as well as content and format of following learning sessions. Thus, session seven was planned to address the forgetting that had most probably occurred over the summer break. The session took place 30 days after session six and focused primarily on reinforcement of the words participants seemed to have more difficulties with.

Although in this session subjects reviewed all of the target words, special attention was paid to the ones they seemed to have most trouble with, as reflected by results from the pre-test and the test in the previous session. Therefore, the researcher

identified the 15 questions participants had mostly answered wrongly to in the pre-test and in session six and collected the target words representing the answer to each one of those questions (see Appendix XV). This resulted in a pool of 17 words (since most of the words in both tests were repeated). Considering some of the target words appeared inflected in the tests (e.g., *hacer* appears as *hago*) but participants had created the flashcards using the base form of target words (e.g., nouns in the singular form, except for *cubiertos*, and verbs in the infinitive form), for consistency, in this activity, target words were presented to students as they appeared in the list in Appendix VIII. The final list of target words for the main activity of the session was:

<i>cena</i> ('dinner')	<i>solicitar</i> ('to request')	<i>viejo</i> ('old')
<i>cubiertos</i> ('silver-ware')	<i>tampoco</i> ('either')	<i>herida</i> ('wound')
<i>regresar</i> ('to return')	<i>caminar</i> ('to walk')	<i>panadería</i> ('bakery')
<i>equipo</i> ('team')	<i>ayudar</i> ('to help')	<i>odiar</i> ('to hate')
<i>llover</i> ('to rain')	<i>soleado</i> ('sunny')	<i>precio</i> ('price')
<i>alquilar</i> ('to rent')	<i>hacer</i> ('to do')	

The session began with subjects receiving a printed list of the 17 Spanish target words listed above. Participants were requested to work in pairs. Each group also received a blank sheet of paper and were asked to spend ten minutes copying the words given on the blank paper. At the same time, they were also asked to retrieve from memory the translation into English of each target word and write them down next to the Spanish equivalent. If they did not remember, they were allowed to look at the online flashcards on Quizlet. Once they had written all of the bilingual word pairs, they were asked to draw a picture portraying the target words.

Later, in order to continue to rehearse all of the 100 target words, and not just the 17 mentioned above, participants were once more asked to refer to their online flashcards on Quizlet for ten minutes. As before, the objective was to work individually and try to retrieve meanings of target words from memory.

For the last activity of the session, participants were divided into groups of three. Participants were asked to provide the appropriate Spanish translation for the English words provided on a purposely-coded interactive activity (see Appendix XVI). In order to help participants find the appropriate translations, hints using scrambled versions of the target words in Spanish were provided.

All students were present and actively engaged in the activities. Finally, in some of the activities participants were allowed to share information with peers, which seemed to show that working in collaboration and in a relaxed environment also seemed to promote concentration, learning, and motivation.

#### **4.4.9 Session eight**

Session eight occurred 31 days after the previous session and it was planned so that participants would be exposed to target words in an implicit manner. The reason behind this idea was to provide a more relaxed environment, without the need to memorize target words, in order to avoid boredom and promote engagement. This was largely based on suggestions from Bahrick (1979) and Schmitt (2008) stating that once vocabulary had been acquired implicit revisits could help retention.

The lesson consisted of three activities, starting with a ‘warm-up’ individual activity followed by pair-review and reflection work. The last and main activity of this session was a friendly competition where participants raced each other in an online interactive race using all target words.

The lesson started by having participants working individually for ten minutes correcting mistakes in ten sentences using target vocabulary (see Appendix XVII) concentrating specially on the 17 words referred to in session seven. Subjects received a printed copy with the ten sentences, and they were asked to focus especially on word meanings. They were also asked to write down a logically correct version of

each sentence on the same paper. After the initial ten-minute activity, participants were instructed to find a partner, and for another ten minutes they had to compare results and agree on one final version per sentence.

The remaining of the session was used for the main activity, which consisted of a space race on Socrative (see Figure 4-5). Participants were asked to work individually and use their personal computers to log in to Socrative where they would find a link to a 'Space race'. The race included all 100 vocabulary items, and participants had to answer multiple choice questions. The questions consisted of the target word in Spanish with three translation options in English. After selecting their response, the next question would appear immediately after until all 100 questions had been answered. If the answer was correct the rocket in the game would move forward, but if the answer was wrong the rocket stayed still. The website offered a separate live view mode where live race status of all participants' rockets were displayed. This was projected on a big classroom screen and subjects could see their own rocket (together with rockets from the rest of the participants) moving forward in race mode towards the finish line as they were answering questions correctly. There was a warm up race at the beginning so participants could understand the game and how live results were displayed on the classroom's big screen. There were three main races, and scores were kept in order for the top three score leaders to get a chocolate bar as a reward.

The session went as planned and students were actively engaged in all three activities. The sentence correction activity helped strengthen target word meaning and to reinforce contextual strategies needed to understand words in context. This was of main importance considering in the pre-test, post-test and delayed post-tests target words appeared in sentences.

Participants seemed to enjoy the online races and loved the friendly competition, specially the rewards. Two students were absent due to an away field trip. They compensated for the missed session upon their return to school (seven days after the session) working under the exact same circumstances with the researcher one day after school.

This session proved to be very successful from three different perspectives. To start with, participants were exposed to all 100 target words, and showed confidence in dealing with them. Contextual strategies were also enhanced in this session, and finally, participants showed great levels of motivation.

#### **4.4.10 Session nine**

This session took place 30 days after session eight, with all participants attending, and it was the last one before the two-week winter break. Therefore, in order to avoid any learning loss that could take place over the break, the aim of this session was to reinforce vocabulary retention through activities that promoted memorization.

Session nine occurred exactly as session two, and it also started with the individual ten-minute revision using the online flashcards (with the retrieval activity). Later, participants were given another ten minutes to revise all 100 target words in any way they preferred. Finally, they were asked to log in to Quia and spend five minutes working on the matching activity, the game, and the translation exercise they had worked on during learning session two. For the remainder of session nine (ten minutes) participants were asked to play any interactive activities using the flashcards on Quizlet.

#### **4.4.11 Session ten**

This session occurred 34 days after the last session, during the first Spanish class after returning from the winter break. Taking into account there was only one session remaining to complete the study, this session consisted mainly of an assessment task. In order to check participant retention of target vocabulary up to this point, students took the same test as in session four (see Appendix XIII) under the same conditions. The test consisted of thirty multiple-choice questions where participants were required to select the right translation from three options. Correct answers were provided at the end of the test as the main goal was to test participants' retention levels and reinforce meaning of words participants might still have problems with. Just as before, after finishing the test, participants were asked to log in to Quizlet and play interactive games using their online flashcards.

#### **4.4.12 Session eleven**

The last learning session took place exactly 30 days after session ten and all subjects were present. The main goal of this lesson was to revise the target words for the last time, and to boost participants' motivation so they would be willing to give their best in the post-tests until the 70-day RI test.

Motivation and commitment were of major importance in the last portion of the study considering there were three remaining post-tests. Session eleven then recycled some of the activities which the researcher considered as successful (meaning that they had enhanced learning and engagement) in previous sessions. The main plan behind this session was to have participants work on activities they had done in the past, so they could contrast their current performance in those exercises and see for themselves how much they had progressed since the beginning of the project. This would ideally enhance confidence that could help keep motivation levels high with the aim that participants would still be motivated and commit to perform at their best until the end of the project.

The actual activities consisted of the usual ten-minute retrieval activity trying to guess the meaning of the flashcards before looking at them. This was followed by another fifteen-minute activity in which participants were given a paper containing all 100 target words in Spanish and they had to provide the English translation or draw a picture. Participants worked in pairs afterwards to check for mistakes. Finally, participants played the space-race activity on Socrative, just as in session eight, under similar circumstances and with chocolate bars as prizes.

With the learning sessions coming to an end, it was time to check actual retention of knowledge. The section below details the three tests taking place after the learning sessions that were used to collect data regarding retention levels.

#### **4.4.13 Post-test (30-day RI)**

Participants in the experimental and the control group took the scheduled post-test 30 days after the last learning session (at 30-day RI). The test was the exact same one participants had taken as a pre-test (see Appendix XII).

All participants were present and went through the exact same procedure as in the pre-test (they took the test online in their usual Ab Initio class time and classroom, and in the presence of their usual teacher that acted as proctor). Before proceeding with the test participants in the experimental group were reminded that they had two scheduled delayed post-tests to take as well. Subjects could use the complete class time (60 minutes) to finish the test and they were allowed to leave the classroom once they had completed the test.

#### **4.4.14 Delayed Post-test (60-day RI)**

Exactly 30 days following the post-test all participants from the experimental group alone took the delayed post-test (at 60-day RI). Test conditions and procedures were exactly the same as in the pre-test and post-test.

#### **4.4.15 Second delayed post-test (70-day RI)**

This test deserves special consideration since, due to circumstances beyond the control of the researcher, it did not take place as planned. Originally this test had been conceived to take place 10 days after the previous test (at 60-day RI) by the experimental group under the exact same conditions as the two previous post-tests.

The test had been scheduled to take place during a time when participants were on study-leave. This meant that subjects were no longer attending classes at the school in preparation for their final exams. Unfortunately, due to several circumstances, not all of the participants managed to take the test on the pre-arranged day. All participants took the test eventually but on three different consecutive days. As a consequence, some participants (three) took the test a day before the pre-arranged day, others (nine) on the same day, and the rest (three) a day after. Finally, considering internet was down when the first group of subjects took the test, the researcher decided that the second delayed post-test would be administered on paper for all participants.

Despite the difficulties mentioned above the test (on all three days) took place without any other distractions, with the researcher acting as proctor. Participants seemed committed and motivated to do their best.

This second delayed test marked the end of the intervention. Test results and analyses are introduced in the following chapter.



## Chapter 5: Results

### 5.1 Introduction

This chapter first presents test results that directly refer to the research questions (defined in 2.3.3 above). All raw scores for the tests can be found in Appendix XVIII. This chapter concludes by analyzing tests that the experimental group alone took during the learning sessions, since their results were fundamental for monitoring participant progress during the learning sessions.

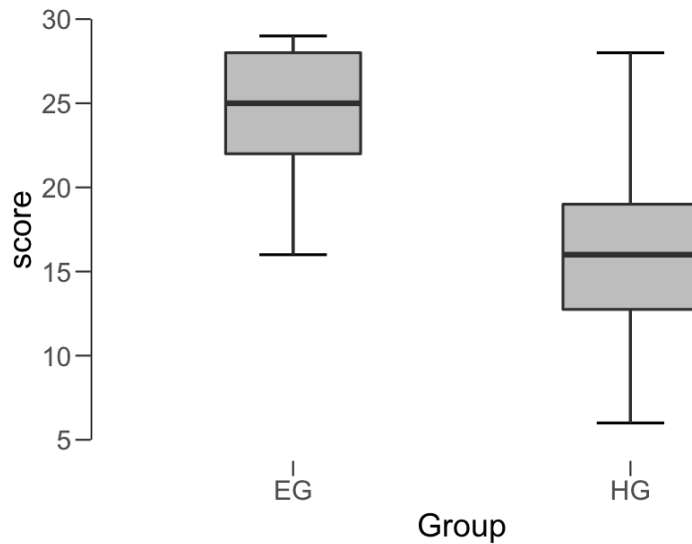
### 5.2 Comparison of test scores of the EG vs. the HG

This section focuses directly on research question one regarding the difference in scores between the experimental group after the intervention, and the group that had graduated a year before (the historical group). Results for this section were obtained from the only test the historical group (HG) took prior to the beginning of the learning sessions, and the first of the post-tests taken by the experimental group (EG) 30 days after the last learning session at 30-day RI.

To analyze the statistical significance of the results, an independent-samples *t*-test was conducted. A first analysis of the data, as shown in Table 5-1 below, reveals that the scores (as represented by their mean) of the experimental group ( $M=24.60$ ) were higher than those of the historical group ( $M=15.84$ ). The experimental group's (EG) results also appeared to be more homogeneous since they had a smaller standard deviation ( $SD=3.869$ ). The results of the historical group (HG) appeared to be more dispersed ( $SD=4.867$ ). Figure 5-1 offers a more graphical representation of the results.

	<b>Group</b>	<b>N</b>	<b>Mean</b>	<b>SD</b>
score	EG	15	24.60	3.869
	HG	32	15.84	4.867

**Table 5-1: EG 30-day RI test vs. HG's test**



**Figure 5-1: EG 30-day RI test vs. HG’s test**

The next step in the analysis was to check whether the samples were normally distributed. A Shapiro-Wilk test (used to check normality distribution) revealed that samples were normally distributed since about 68 percent of the data values were within one standard deviation below and above the mean. Results of the Shapiro-Wilk test are shown in Table 5-2 below where the P values for both groups were not statistically significant (EG=0.084 and HG=0.704) as  $p > 0.05$ . The test estimated that the variance of the sample was normally distributed as the P value returned was greater than 0.05. It is important to note at this time, that this thesis accepts a significance level of 0.05 ( $\alpha=0.05$ ).

		<b>W</b>	<b>p</b>
score	EG	0.896	0.084
	HG	0.977	0.704

**Table 5-2: Test of Normality (Shapiro-Wilk)**

The final step in the *t*-test process was to run a Students' *t*-test (a parametric test) that returned the values shown in Table 5-3. Results show a *t* value of 6.110 which resulted from measuring the size of the difference relative to the degrees of freedom (df=45). This means that on average the experimental group (M=24.60, SD=3.869) scored significantly higher than the historical group (M=15.84, SD=4.867),  $t(45)=6.110, p<.001, d=1.912$ . The difference then between groups is statistically significant (as  $p<.05$ ). The same table also shows Cohen's *d* (d) scores, which is a measure of effect size between the two means and analyzes the practical usefulness of the findings. According to Cohen's conventions, the effect size is said to be small if the value is from 0.2 to 0.5, it is medium if values run from 0.5 to 0.8, and values are said to be large if over 0.8. Based on this, Table 5-3 shows that there was a very large size effect between the experimental group (EG) and the historical group (HG).

	<b>t</b>	<b>df</b>	<b>p</b>	<b>Cohen's d</b>
score	6.110	45.00	< .001	1.912

**Table 5-3: Independent samples Students' *t*-test (EG vs. HG)**

This section revealed that (as shown in Table 5-1) there was a very clear difference in scores between the experimental group (EG) and the historical group (HG), and that this difference was highly statistically significant with a very large effect size. The following section will compare results from the experimental group (EG) and the control group (CG) in a similar fashion.

### **5.3 Comparison of post-test scores of the EG vs. the CG**

Before analyzing results of the post-test of the experimental group (EG) and the control group (CG), this section begins by contrasting results of both groups in their pre-test as this data will be later used in the Discussion chapter to compare results between the groups at different points in time.

Another independent-samples *t*-test was run to analyze results of these test scores. A Shapiro-Wilk normality check revealed that samples were normally distributed at  $p=0.704$  for the control group (CG) and  $p=0.648$  for the experimental group (EG) (see Table 5-5), therefore, a Student's *t*-test was run to contrast results. Findings below (Table 5-4 and Table 5-6) show that prior to the intervention there was no significant difference between the experimental group ( $M=13.47$ ,  $SD=2.615$ ) and the control group ( $M=12.17$ ,  $SD=3.512$ ),  $t(25.00)=-1.103$ ,  $p=0.280$ . A Cohen's *d* value of  $-0.427$  shows a small size effect. It is important to notice that the negative values shown by Cohen's *d* and *t* in the table, simply have to do with the fact that when the test was run the group with the smallest mean was listed first, therefore, the results are negative. This does not affect the significance of the results. Figure 5-2 shows a boxplot which graphically displays groups' scores prior to the intervention.

	<b>Group</b>	<b>N</b>	<b>Mean</b>	<b>SD</b>
Score	CG	12	12.17	3.512
	EG	15	13.47	2.615

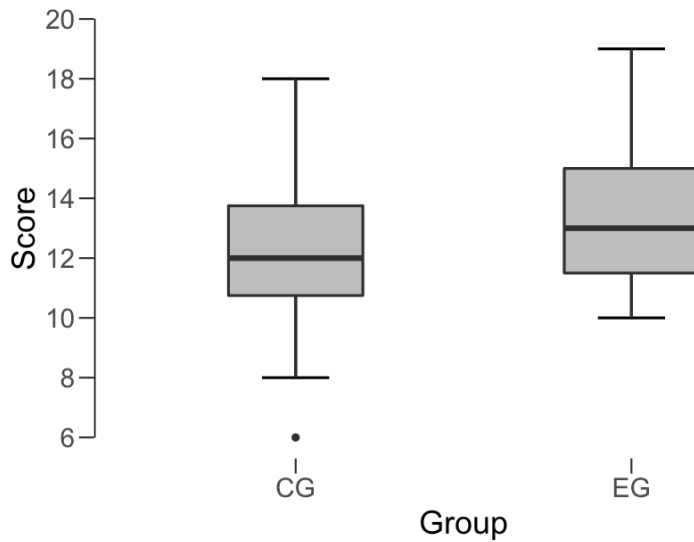
**Table 5-4: Pre-test EG vs. CG**

		<b>W</b>	<b>p</b>
Score	CG	0.955	0.704
	EG	0.957	0.648

**Table 5-5: Test of Normality (Shapiro-Wilk) for pre-tests**

	<b>t</b>	<b>df</b>	<b>p</b>	<b>Cohen's d</b>
Score	-1.103	25.00	0.280	-0.427

**Table 5-6: Independent samples *t*-test EG vs. CG**



**Figure 5-2: Pre-test EG vs. CG**

After inspecting the difference between both groups' pre-test, the results of their post-test were also analyzed. In order to evaluate the statistical significance of the results, another independent samples *t*-test was conducted. A Student's *t*-test was used since a Shapiro-Wilk test revealed that results from both samples were normally distributed as results were not significant at  $p=0.424$  for the control group (CG) and  $p=0.084$  for experimental group (EG) (see Table 5-7).

Both Table 5-8 and Table 5-9 below show that on average participants in the experimental group ( $M=24.60$ ,  $SD=3.869$ ) scored significantly higher than the control group ( $M=17.00$ ,  $SD=4.553$ ),  $t(25.00)=-4.690$ ,  $p<.001$ ,  $d=-1.817$ . A  $p<.05$  shows that the difference between groups is statistically significant and there is also a very large effect size.

		<b>W</b>	<b>p</b>
score	CG	0.934	0.424
	EG	0.896	0.084

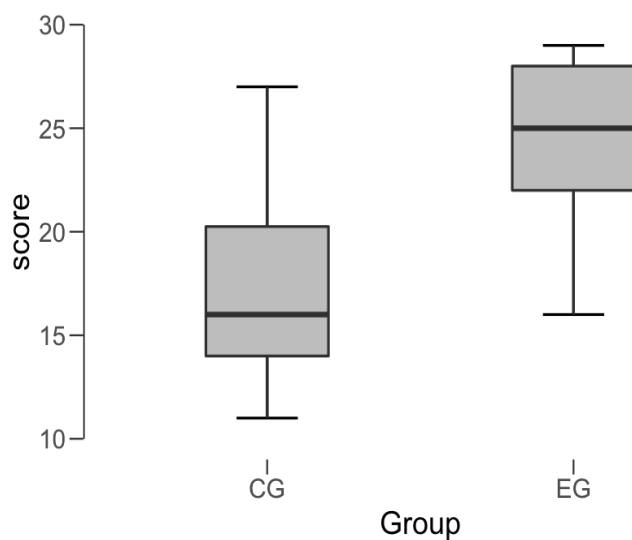
**Table 5-7: Test of Normality (Shapiro-Wilk) for post-tests**

	<b>Group</b>	<b>N</b>	<b>Mean</b>	<b>SD</b>
score	CG	12	17.00	4.553
	EG	15	24.60	3.869

**Table 5-8: EG 30-day RI test vs. CG's post-test**

	<b>t</b>	<b>df</b>	<b>p</b>	<b>Cohen's d</b>
score	-4.690	25.00	< .001	-1.817

**Table 5-9: Independent Samples Students' *t*-Test (EG vs. CG)**



**Figure 5-3: EG 30-day RI test vs. CG's post-test**

Figure 5-3 clearly shows the difference between the groups, and how much more dispersed results in the control group were in comparison to those of the experimental group's.

The previous analyses compared results between the experimental group and the historical group, and the experimental group versus the control group (EG vs. HG, and EG vs. CG). In contrast, the following sections will seek to compare results of the different tests taken by the experimental group at different points in time in order to analyze the group's progress.

#### **5.4 Experimental group's pre-test, 30-day RI and 60-day RI post-tests**

This section analyzes test results of the experimental group in order to see its progression starting from the pre-test until the 60-day RI post-test. The main goal of this section is to inspect data that could help answer research question three that investigated which retention interval obtained the highest retention scores, given the project's 30-day ISI (see research question three in 2.3.3 above).

In order to provide a complete picture of the experimental group's performance, this section would ideally need to consider results of the pre-test, plus each one of the three post-tests (30-day, 60-day, and 70-day RI). As explained in 4.4.15 above, the last post-test (70-day RI) did not take place as expected. As a consequence, only the pre-test and the first two post-tests (30-day and 60-day RI) were considered in this section. Results and a thorough analysis of the 70-day RI test will be discussed in the following section.

The first comparison between the tests can be seen in Table 5-10. The table reveals that the scores of the 30-day RI test are higher than the other two ( $M=24.60$ ), although there is only a slight difference in comparison to the 60-day RI test ( $M=24.13$ ). Results of the pre-test are the most homogeneous since they have the smallest standard deviation ( $SD=2.615$ ). This difference in results can also be appreciated in Figure 5-4 further below. In order to investigate whether this apparent difference between the tests was statistically significant (since the samples came from the same group) a repeated-measures ANOVA test was conducted. The test was used to calculate  $F$  which represents the variance between the samples. An  $F$  value, together with a sample size (as expressed by the degrees of freedom =  $df$ ) provide the necessary information to calculate the value of  $P$ , which in place will serve to express the significance of the results. The repeated-measures ANOVA test requires that the

equality of the variances between the samples should be met. Therefore, a Mauchly's Test of Sphericity was run. Table 5-11 shows results of such test which presents the value of P as 0.547. This means that there was no significant difference in the variances as  $p > 0.05$  and therefore the ANOVA test could be run without applying any sphericity corrections to it.

<b>Retention</b>	<b>N</b>	<b>Mean</b>	<b>SD</b>
Pre-test	15	13.47	2.615
30-day RI	15	24.60	3.869
60-day RI	15	24.13	4.068

**Table 5-10: EG's pre-test, 30-day and 60-day RI post-tests**

	<b>Mauchly's W</b>	<b>p</b>	<b>Greenhouse-Geisser <math>\epsilon</math></b>	<b>Huynh-Feldt <math>\epsilon</math></b>
Retention	0.911	0.547	0.919	1.000

**Table 5-11: Test of Sphericity**

The ANOVA test (see Table 5-12 below) shows that combined results of all tests were statistically significant  $f(2,28)=165.0, p < .001$ . These results revealed that a significant difference existed between groups, but in order to know exactly where the differences were (between the groups), a *post-hoc* test was also run. In this case a Bonferroni test was preferred as it seems to work well for small sample sizes lower than 30 (Lowie & Seton, 2012). The *post-hoc* analysis (see Table 5-13) revealed that both the difference between pre-test ( $M=13.47, SD=2.615$ ) and the 30-day RI test ( $M=24.60, SD=3.869$ ), and the difference between pre-test and the 60-day RI test ( $M=24.13, SD=4.068$ ) were significant at  $p < 0.001$ . The difference between the 30-day RI test and the 60-day RI test was not significant as  $p=1.000$ . The same table also shows Cohen's d (d) scores which reveals large effect sizes in results contrasting the pre-test and the 30-day RI test, and the pre-test and the 60-day RI test. However, there is a small size effect when contrasting the 30-day and the 60-day RI tests.



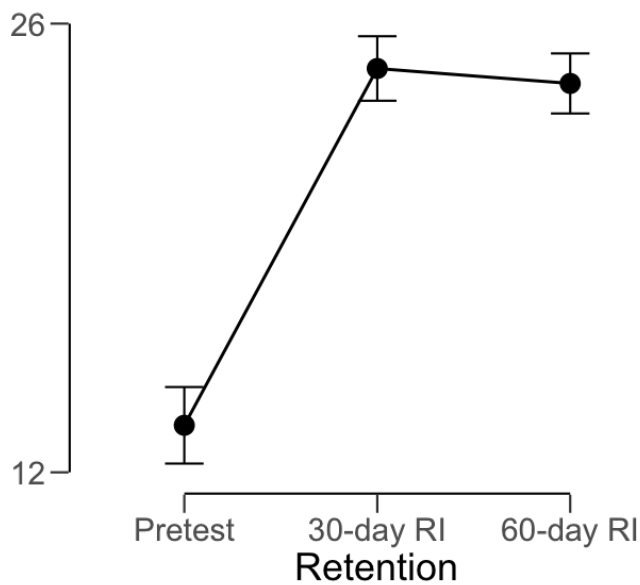
	Sum of Squares	df	Mean Square	F	p
Retention	1189.7	2	594.867	165.0	< .001
Residual	100.9	28	3.605		

**Table 5-12: ANOVA test results of pre-test, 30-day RI and 60-day RI tests**

		Mean Difference	t	Cohen's d	p <sub>bonf</sub>
Pre-test	30-day RI	-11.133	-14.623	-3.776	< .001
	60-day RI	-10.667	-14.783	-3.817	< .001
30-day RI	60-day RI	0.467	0.798	0.206	1.000

**Table 5-13: Post Hoc Comparisons - Retention**

As seen in the descriptive plot below, results from the 30-day RI test were the highest of the three tests analyzed. It is easy to see also, how results after the intervention were comparatively higher in relation to those prior to the treatment.



**Figure 5-4: EG's pre-test, 30-day and 60-day RI post-tests (out 30 possible points)**

This section served to contrast results of three of the tests taken by the experimental group. The remaining test (70-day RI post-test) could not be easily contrasted with the rest of the tests here and is described below.

### **5.5 70-day RI post-test**

As explained in 4.4.15 above, the last test of the project did not take place as planned (the test was on paper, not on a computer, and some participants took it on the planned day, but some others took it a day before, or a day after the set day). Considering this could have affected participants' outcomes (for instance, some participants may do better in tests on paper than on a computer, or an extra retention day, or one day less could have affected retention overall) results from the last post-test were not very reliable, so they were not used in a straight forward comparison against tests in the previous section.

At the same time, however, considering results of this test still show relevant vocabulary retention information, it seemed important to analyze these results, with caution, and contrast them against the other tests. Therefore, the analysis below expands the findings from the previous sections by displaying information of all tests but concentrating specially on those of the 70-day RI test.

To start with, Table 5-14 below shows descriptive statistics of all three test conditions. A test of Sphericity (Table 5-15) showed that, as opposed to the previous section, there was significant variance between the groups ( $p < .05$ ). Therefore, statistical correction of results was needed (these results are reflected in Table 5-16).

<b>Retention</b>	<b>N</b>	<b>Mean</b>	<b>SD</b>
Pre-test	15	13.47	2.615
30-day	15	24.60	3.869
60-day	15	24.13	4.068
70-day	15	22.40	4.154

**Table 5-14: EG's pre-test, 30-day, 60-day, and 70-day RI post-tests**

	<b>Mauchly's W</b>	<b>p</b>	<b>Greenhouse-Geisser <math>\epsilon</math></b>	<b>Huynh-Feldt <math>\epsilon</math></b>
Retention	0.104	< .001	0.667	0.779

**Table 5-15: Test of Sphericity for 70-day RI test**

The results of the ANOVA test with sphericity corrections (see Table 5-16) show that there was still an overall significant effect on vocabulary retention considering all tests together:  $f(2.002, 28.025)=134.8, p<.001$ . The particular difference between the 70-day RI test and the rest of the tests was obtained through a *post-hoc* Bonferroni test (Table 5-17). The *post-hoc* analysis revealed that there was a statistically significant difference at  $p<.05$  between the 70-day RI post-test ( $M=22.40, SD=4.154$ ) and the rest of the tests: pre-test ( $M=13.47, SD=2.615$ ), 30-day RI test ( $M=24.60, SD=3.869$ ), and 60-day RI test ( $M=24.13, SD=4.068$ ). Finally, results show large effect sizes among the 70-day RI post-test and all of the other tests.

	<b>Sphericity Correction</b>	<b>Sum of Squares</b>	<b>df</b>	<b>Mean Square</b>	<b>F</b>	<b>p</b>
Retention	Greenhouse-Geisser	1221.0	<sup>a</sup> 2.002	<sup>a</sup> 609.957	<sup>a</sup> 134.8	<sup>a</sup> < .001
Residual	Greenhouse-Geisser	126.8	28.025	4.523		

**Table 5-16: Within subjects effects for 70-day RI post-test**

		<b>Mean Difference</b>	<b>SE</b>	<b>t</b>	<b>Cohen's d</b>	<b>p<sub>bonf</sub></b>
Pre-test	30-day	-11.133	0.761	-14.623	-3.776	< .001
	60-day	-10.667	0.722	-14.783	-3.817	< .001
	70-day	-8.933	0.733	-12.182	-3.145	< .001
30-day	60-day	0.467	0.584	0.798	0.206	1.000
	70-day	2.200	0.641	3.430	0.886	0.024
60-day	70-day	1.733	0.153	11.309	2.920	< .001

**Table 5-17: Post Hoc Comparisons - 70-day RI post-test**

To conclude, the boxplot below (Figure 5-5) shows that the 70-day RI test had overall higher results in comparison to the pre-test, and it was also the one with the lowest results of three retention tests. The same figure reveals that, overall, participants in the experimental group obtained the highest vocabulary retention scores 30 days after the last learning session.

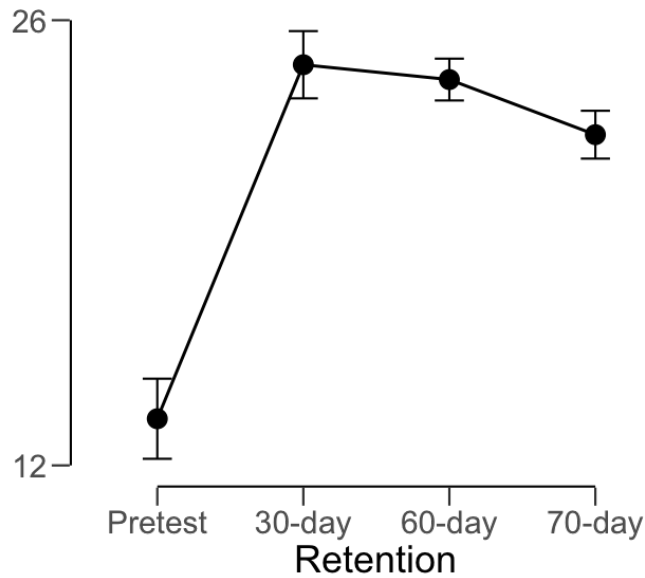


Figure 5-5: EG's pre-test, 30-day, 60-day, 70-day RI post-tests (out 30 possible points)

## 5.6 Results of the tests within the learning sessions

The tests described below were used to monitor student learning at a certain point during the treatment (in session four, six, and ten). Data obtained from those tests was used to assess subject learning while the project was in progress, and to adjust preplanned-learning sessions if necessary.

### 5.6.1 Tests in session four and ten

This section introduces results of the test administered in session four and ten together considering the test was the exact same one (administered twice) making it easier to analyze the progress of the experimental group during the treatment. This allows to see individual test scores, while it also contrasts results of both tests to inspect participant learning while the project was still ongoing.

Considering samples came from the same group, a paired-samples *t*-test was run to analyze results. A Shapiro-Wilk normality check revealed that samples were normally distributed at  $p=0.076$  (see Table 5-19), therefore, a Student's *t*-test was run to contrast results. Findings below (see Table 5-18, and Table 5-20), show that there was

a significant difference between them, and results of the test in session ten (M=48.67, SD=2.093) were higher and more homogeneous than in session four (M=45.73, SD=4.431),  $t(14)=-3.093$ ,  $p=.008$ . Finally, a Cohen's d value of -0.798 shows a medium size effect.

	<b>N</b>	<b>Mean</b>	<b>SD</b>
session 4	15	45.73	4.431
session 10	15	48.67	2.093

**Table 5-18: Results of the tests in session four and ten**

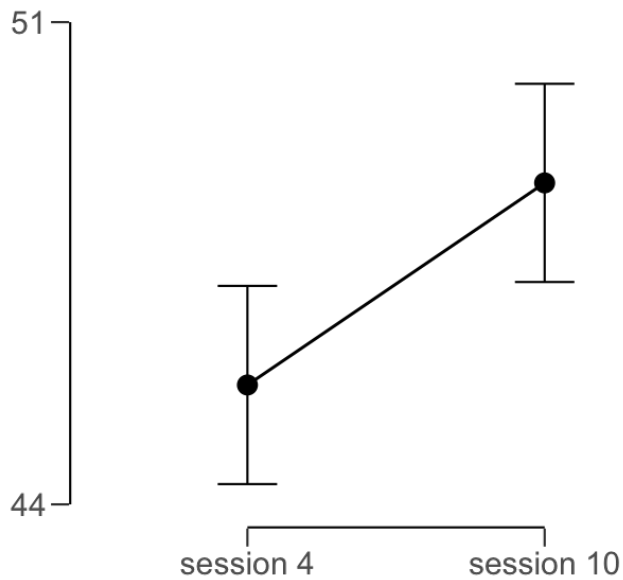
	<b>W</b>	<b>p</b>
session 4 - session 10	0.894	0.076

**Table 5-19: Test of Normality (Shapiro-Wilk) session 4 and 10**

	<b>t</b>	<b>df</b>	<b>p</b>	<b>Cohen's d</b>
session 4 - session 10	-3.093	14	0.008	-0.798

**Table 5-20: Paired samples t-test session 4 and 10**

Figure 5-6 below shows that, with a maximum possible test score of 50 points, in session four, a general mean value of 45.733 indicates that overall percentage scores (91.5%) were very high. The same figure shows that test scores in session ten were even higher with a mean value of 48.667 and a group average of 97.3%.



**Figure 5-6: Boxplot of results of the tests in session four and ten**

### 5.6.2 Test in session six

The test in session six (see 4.4.7) was an exact replication of the pre-test and it took place 90 days after the previous learning session. This section compares both tests (pre-test and session six) to analyze the progression of the experimental group during the learning sessions. The analysis arising from this section will be further inspected in the Discussion chapter to investigate the overall experimental group's performance in the study.

As in the previous section, since samples came from the same group, a paired-samples *t*-test was run to analyze results of these two tests. A Shapiro-Wilk normality check revealed that samples were normally distributed at  $p=0.240$  (see Table 5-22), therefore, a Student's *t*-test was run to contrast results. Test scores (see Table 5-21 and Table 5-23 below) showed that, on average participants performed significantly better in session six ( $M=18.20$ ,  $SD=5.335$ ) in comparison to the pre-test ( $M=13.47$ ,  $SD=2.615$ ),  $t(14)=-5.071$ ,  $p<.001$ . The lower SD in the pre-test reflects that scores were more homogeneously distributed among participants prior to the study. In contrast, the higher SD in session six shows that test results were more dispersed. Finally, a Cohen's *d* value of  $-1.309$  shows a large size effect.

	<b>N</b>	<b>Mean</b>	<b>SD</b>
Pre-test	15	13.47	2.615
Session 6	15	18.20	5.335

**Table 5-21: Pre-test & session-six test mean values**

	<b>W</b>	<b>p</b>
Pre-test - Session 6	0.926	0.240

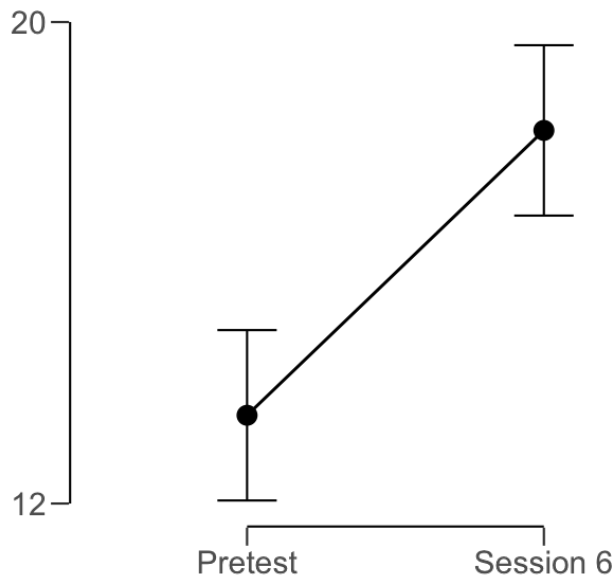
**Table 5-22: Test of normality (Shapiro-Wilk) test in session six**

	<b>t</b>	<b>df</b>	<b>p</b>	<b>Cohen's d</b>
Pre-test - Session 6	-5.071	14	< .001	-1.309

**Table 5-23: Pre-test & session-six t-test**

The figure below (see Figure 5-7) also shows the difference in learning between the two tests. Out of 30 possible points in the pre-test there was a general group mean value of 13.467, while in session six, the group mean score was 18.2.





**Figure 5-7: Pre-test & session-six test mean values**

Finally, tests' scores revealed that there had been a group net increase of 15.8% comparing the percentages in both pre-test and the test in session six. This showed that half way through the study overall participants' scores had reached 60.7% after five learning sessions and a 90-day RI gap.

Results presented in this chapter will be discussed in the following chapter. All results' interpretation and how they relate to the main goals of this project are introduced next.

## Chapter 6: Discussion

### 6.1 Introduction

The replication study and the main research project of this thesis explored the efficacy of spaced repetition to improve L2 vocabulary retention. Findings from those studies are consistent with previous research suggesting that spaced repetition could indeed enhance vocabulary retention. Despite those results the actual implementation of spaced repetition (SR) in a real language course may not be so straightforward after all. This chapter will interpret those findings and contrast them against previous research in an attempt to move forward towards an optimal implementation of spaced repetition in actual educational settings. Limitations of this thesis and potential areas for future research are presented at the end of the chapter.

The research study introduced in chapter three consisted of a replication study of the original project by Johnson & Heffernan (2006) that employed spaced repetition to help language beginners comprehend authentic materials (movie trailers). In the replication study 50 students of L2 Spanish received spaced repetition instruction to learn and practice an average of ten target words per reading in a series of seven readings based on seven movie trailers. In a pre/post-test design after three learning sessions participants appeared to increase their knowledge of target vocabulary and seemed to improve their understanding of authentic materials. In comparison to similar previous research (see Table 3-11), with a post-test occurring immediately after the last learning session (at 0-day RI) the original study and the replication were among the lowest scoring studies in the table. Johnson & Heffernan (2006) and the replication study were the only ones with interstudy intervals (ISI) longer than one day (the original study applied a 7-day ISI across nine learning sessions and the replication study had three learning sessions at 2/5-day ISI). The fact that the original and the replication study scored lower than previous studies could be attributed to the fact that learning seems to be faster at shorter interstudy intervals (ISI) but information is forgotten more quickly. Longer interstudy intervals, on the other hand, seem to show slow-paced learning, but information appeared to be remembered longer than through shorter interstudy intervals. Another reason for low retention

scores in both studies could be attributed to the fact that the number of rehearsals was rather low, since more rehearsals enhance learning and retention. Finally, the replication study, similar to the original study, could not produce proper retention data necessary to gauge vocabulary retention after rehearsals had stopped. Also, similar to the original study participant motivation seemed to negatively affect study results.

The main research project of this thesis was a longitudinal ecological study that had high-school language learners as participants. The experimental group learned 100 Spanish target words through a spaced repetition treatment for eleven learning sessions at 30-day ISI. Retention was tested at different intervals (30-day, 60-day and 70-day RI) with the 30-day RI post-test showing the highest retention results with an overall average percentage score of 82%. The experimental group obtained the highest percentage score in comparison to the two other groups in the study. Finally, in contrast to the replication study, participant engagement and motivation was very high all across the main research project. Results from the main study revealed that spaced repetition enhanced learning and retention of vocabulary, and that highest retention gains were obtained when the retention interval and the interstudy interval were similar in length.

## **6.2 SR as a means to enhance vocabulary retention**

Considering spaced repetition (SR) had long been claimed to promote vocabulary retention (e.g., Thorndike, 1908; Bahrick, 1979; Bahrick & Phelps, 1987; Dempster, 1989; Cepeda et al., 2006; Ebbinghaus, 2013) but since it was still not being implemented in everyday instructional settings, several scholars (e.g., Bloom & Shuell, 1981; Sobel et al. 2011; Goossens et al., 2012; Küpper-Tetzel et al., 2014; Gryzelius, 2016; and Lotfolahi & Salehi, 2017) conducted spaced repetition (SR) research focusing mainly on field studies contrasting it against massed repetition (MR). While this highlighted the benefits of spaced repetition over massed repetition for long term retention the actual everyday teaching methodologies seemed to be missing. Arguably, in order to test whether spaced repetition can actually bring benefits to a language course, then, the best comparison seems to be to contrast spaced repetition against what actually happens in the every-day classroom. This thesis, then set out with the aim of assessing that and suggested a field study where

spaced repetition is contrasted against a traditional teaching method (i.e., the actual teaching strategies in a language course that takes place daily in a real high-school environment).

With that in mind two first research questions arose in this study:

1. Will spaced repetition produce different retention levels in participants in the experimental group in comparison to students who were taught using traditional teaching methods and graduated a year earlier?
2. Will spaced repetition produce different retention levels in comparison to a control group who were taught using traditional teaching methods?

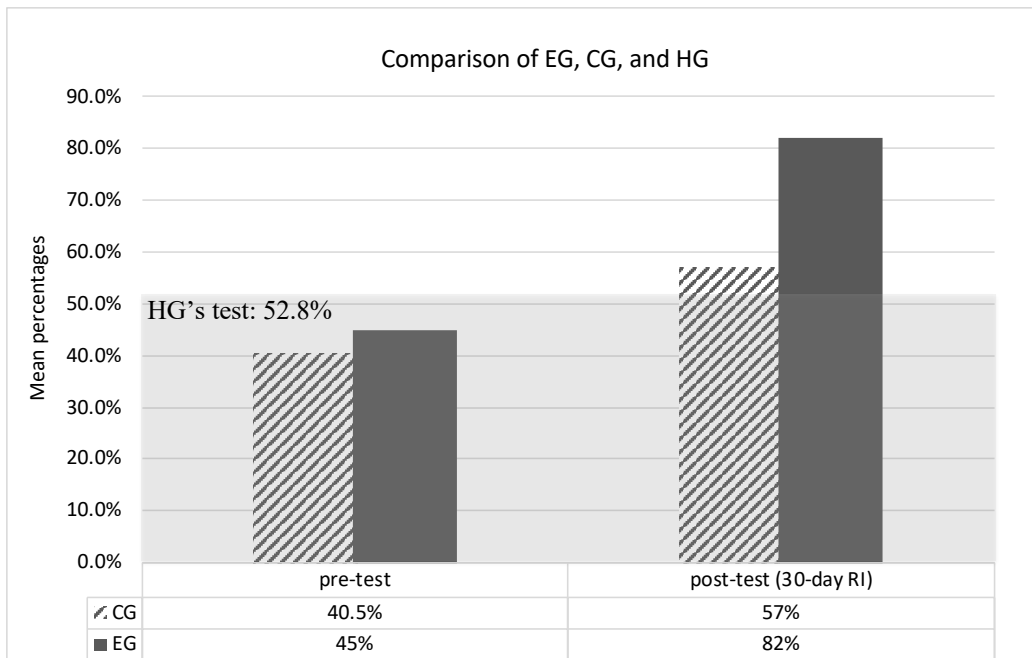
Although relatively similar, these two research questions examined whether an experimental group undergoing a spaced repetition treatment could retain more target words than two groups that did not receive any spaced repetition treatment. The first research question of the project investigated how the 30-day RI post-test scores of the experimental group (EG) compared against those of the historical group (HG). The question actually investigated whether, after eleven learning sessions, spaced repetition (SR) instruction would produce different retention scores (at 30-day RI) against historical data that would be representative of a typical IB Spanish Ab Initio course.

Research question number two examined test score results of the experimental group (EG) vs. the control group (CG). The difference between this comparison and the one in the previous paragraph is that this time the groups were an excellent match to contrast their performance since they were very similar in every aspect (both groups had started the Spanish Ab Initio course at the same time, were at a very similar level when they took the pre-test as shown in 5.3 above, and had almost the same number of participants).

Chapter 5 above provided the results of all the different tests and revealed a clear picture of how the groups compared in relation to one another in terms of knowledge

levels by the end of the study. Figure 6-1 below shows that the historical group (represented by the shaded area) had obtained an overall score of 52.8% in their test. After the learning sessions the first of the post-tests that the experimental group took (at 30-day RI) revealed a considerable difference against the historical group. This time, the experimental group obtained an overall average score of 82%, which unveiled a substantial difference between both conditions in favor of the group that had received spaced repetition instruction (EG).

In relation to the control group, on the other hand, the same figure below shows that (according to their pre-test average scores: EG=45% and CG=40.5%) prior to the learning sessions both groups were almost equal in terms of Spanish knowledge, or at least in their knowledge of the target words. The post-test, however, brought more revealing results. After the intervention, and after a 30-day retention interval the control group scored 57% while the experimental group obtained an overall percentage score of 82%. This highlighted the fact that the experimental group outscored the control group by 4.5% in the pre-test, but that difference increased to 25% after the learning sessions. By looking at Figure 6-1 again, it is interesting to see that in the post-test the control group scored just slightly higher than the historical group which shows that the control group was most probably within the level of Spanish expected to be reached by the end of the course. This reinforces the notion that the spaced repetition condition obtained higher retention gains not only against a very similar group (CG), but that distinction could also be expanded to a larger sample that was representative of the level of knowledge students would generally have at the end of their Spanish language course every year.



**Figure 6-1: Test result comparison of the three groups (EG, CG and HG)**

The fact that in this study participants under the spaced repetition condition scored higher in retention tests than the rest aligns with findings from several scholars who reported that groups under spaced repetition condition obtained higher retention scores in comparison to other groups. For instance, Bahrick (1979) found that participants under a spaced repetition condition outscored those under a massed repetition condition, specially at 30-day RI after rehearsing every 30 days. This is of particular relevance considering Bahrick also taught and tested participants in their knowledge of Spanish words and because one of the spaced repetition conditions was similar to the project in this thesis (30-day ISI and 30-day RI). It is interesting to note however, that while Bahrick conducted a lab study with university students, this thesis reported similar findings in a field study with high-schoolers as participants. Similar outcomes were also found in other articles mentioned in the same chapter. For instance, Bloom & Shuell (1981), Küpper-Tetzel et al. (2014) and Gryzelius, (2016) also reported that spaced repetition seemed to benefit long-term vocabulary retention in foreign language courses with middle-school and high-school learners. The same

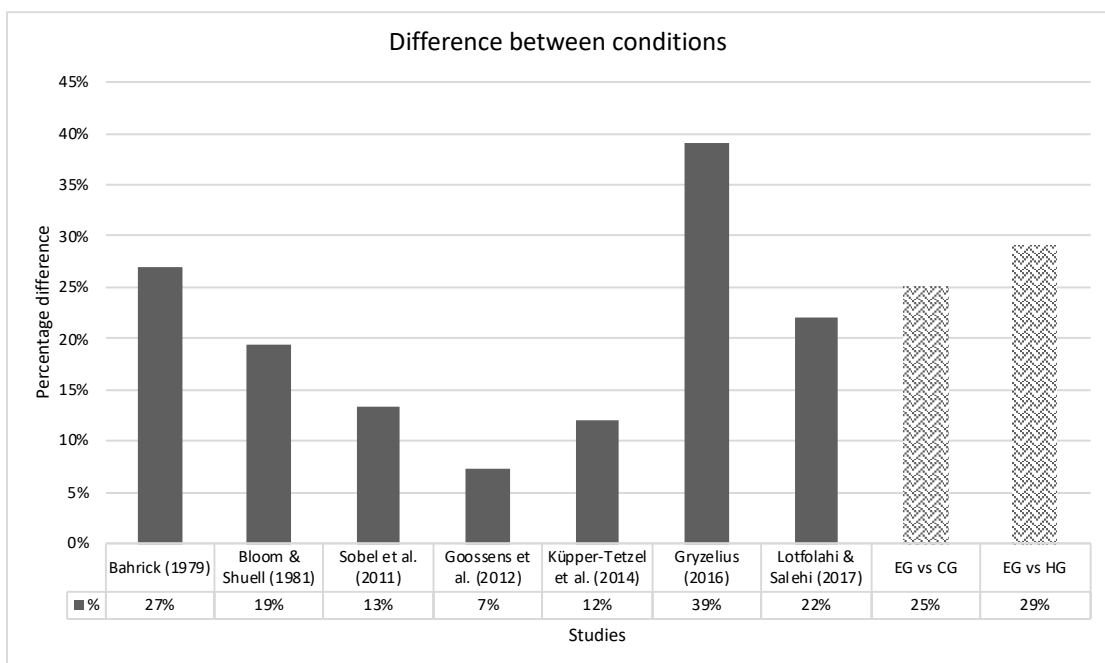
was true for Lotfolahi & Salehi (2017) who reported similar results but having primary school children as participants. The benefits of spaced repetition were also reported in L1 studies as reported by Sobel et al. (2011) and Goossens et al. (2012).

While the main study in this thesis supported those findings, it also expanded on those considering it introduced a field study that was longer than any of the ones mentioned in the previous paragraph. For instance, in this field study the learning sessions alone extended for eleven months, which is only comparable to the largest study presented in Bahrick (1979) which extended for six months. The longest field study of those mentioned above was (in terms of learning phase) Küpper-Tetzel et al. (2014) which extended only for 20 days. Findings from the main research project of this thesis showed then that in a field study, high-school learners of L2 Spanish under a spaced repetition condition (when tested at 30-day RI) retained (as a group average) 82 target words (out of 100) after eleven learning sessions at a 30-day ISI.

The difference between the conditions (spaced repetition vs. traditional teaching methods) in this research project was also contrasted against previous research. Figure 6-2 below offers a clear comparison between the results from this thesis and the literature (notice that only retention intervals that were more closely related to this thesis were selected from previous research). Except for the research project of this thesis, all of the studies mentioned in Figure 6-2 contrasted spaced repetition (SR) against massed repetition (MR). The same figure shows that in this thesis at 30-day RI, the difference in average scores between the experimental group and the control group was 25%, while the difference between the experimental group and the historical group was 29.2% (notice that for easier comparison the difference between the experimental group and the historical group has been rounded to 29%). This is particularly salient considering that Cepeda et al. (2006) reported an average of 15% difference between space repetition (SR) and massed repetition (MR) in a long review of the literature after analyzing 184 spaced repetition studies.

Regarding the most salient studies in Figure 6-2, the study that reported the largest difference between conditions was Gryzelius (2016). The 39% difference (between spaced repetition and massed repetition) in the study could be attributed to the fact

that participants in the massed repetition condition rehearsed all 20 target words for 30 minutes once only and were tested 39 days later. This could mean that time of instruction was most probably too short for the given retention interval. Similar findings are found in Bahrck (1979) with six daily learning sessions it seems that the quick cumulative learning of massed repetition could not survive a 30-day retention interval. Finally, findings from this thesis are below Bahrck's (1979), but above the rest of the studies probably suggesting that the repetition used by the traditional teaching method at the school (see 1.3), probably helped with vocabulary retention, but was not strong enough.



**Figure 6-2: Difference between SR vs other conditions across studies**

To conclude, this study contributed to the field by providing further evidence regarding the role of spaced repetition in long-term vocabulary retention. This means that spaced repetition benefits appear to be consistent across participant age, lab study findings still seem to hold in ecological environments, and spaced repetition appears to produce higher retention gains than both massed repetition conditions and



traditional teaching methods. Results from the main project of this thesis revealed that the spaced repetition condition retained more vocabulary than both the control and the historical group. The combination of a 30-day ISI and a 30-day RI together with teaching strategies to enhance acquisition seemed to foster long-term vocabulary retention.

### **6.3 Optimal RI/ISI combination**

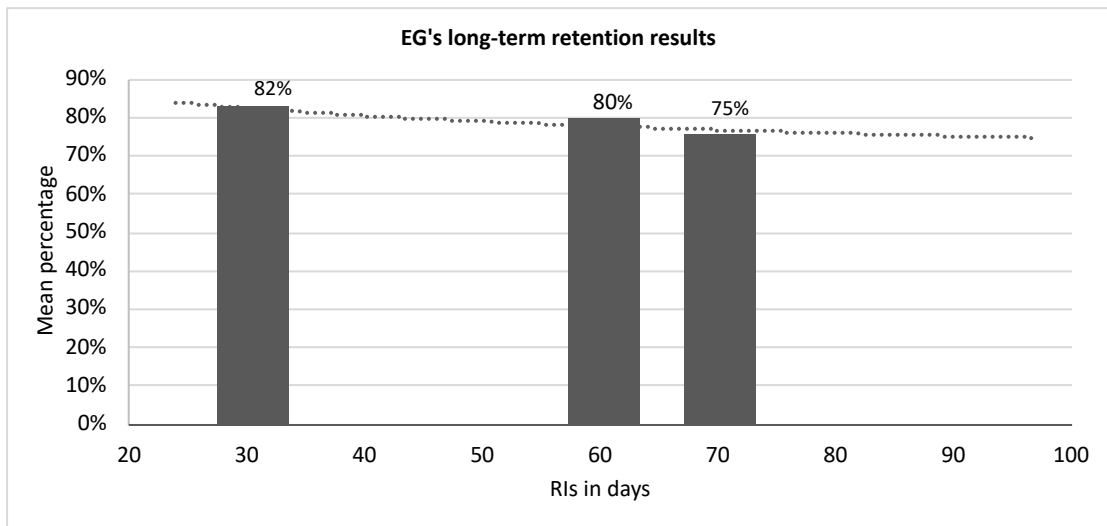
The third research question of this thesis (which as discussed in 2.3.3 was more controversial) sought to determine the optimal retention interval (RI) given a certain lag between learning sessions (interstudy interval). This section then will analyze results that directly address that RQ3:

Given a 30-day ISI, which RI provides the highest retention scores: 30-day, 60-day, or 70-day RI?

Several scholars stated that for long-term vocabulary retention, the retention interval (RI) should be determined first, and then plan the interstudy intervals (ISI) accordingly (e.g., Bahrick, 1979; Cepeda et al., 2006; Cepeda et al., 2008). This however is a controversial matter since there is still uncertainty regarding the optimal RI/ISI combination as several studies (e.g., Cepeda et al., 2006; Rohrer & Pashler, 2007; Cepeda, et al., 2008 and Küpper-Tetzl, et al., 2014; and more recently, Suzuki & DeKeyser, 2017 and Serrano & Huang, 2018) found that the interstudy interval (ISI) should be only a portion of the retention interval (RI). Unlike those studies, Bahrick (1979) and Lotfolahi & Salehi (2017) however, argued that the interstudy interval and the retention interval should be equal in length.

The results of this thesis (see Figure 6-3) show that the highest retention scores were obtained at 30-day RI, dropping towards longer retention intervals. This means that at least in this study the optimal retention interval for the selected interstudy interval was 30 days. This therefore corroborates the idea that the interstudy intervals (ISI) should be equal in length to the retention interval (RI) in order to obtain the highest retention

gains (Bahrick, 1979 and Lotfolahi & Salehi, 2017). It is also true that an earlier post-test (e.g., 10-day or a 20-day RI post-test) could have brought more revealing results. For example, an earlier post-test could have provided more data to see whether test results were higher or lower than the 30-day RI test results.



**Figure 6-3: EG's long-term retention results with trendline**

In Figure 6-3 it is surprising to see that even when 30-day RI appeared to be the optimal retention interval (RI) for an ideal interstudy interval (ISI) of 30 days, at 60-day RI, retention results were still very similar to the 30-day post-test (only 2% less). This seems to imply that the optimal RI/ISI combination might not necessarily be an exact number. A 30-day ISI might produce similar retention results at 30-day RI and also a few days more towards somewhere between 30-day and 60-day RI. The same might also be true a few days before the 30-day RI. This actually is consistent with the lack of agreement found in the literature regarding the best RI/ISI combination. Cepeda et al. (2008) and Küpper-Tetzel et al. (2014), for example, agreed on the notion that the interstudy interval should be a portion of the retention interval, but they still did not agree on an exact value. Also, in a very long review of the literature, Cepeda et al. (2006) reported a very large range among studies regarding the ideal interstudy interval for a given retention interval. For instance, the article reported that

for a retention interval of 30 to 2,900 days, scholars were reporting ideal interstudy intervals between 29 to 168 days in length. Arguably, the explanation of the findings in this thesis could be that vocabulary knowledge might not decay so quickly after rehearsals have stopped after all, or that an optimal combination of ISI and RI together with strong encoding could result in longer and higher vocabulary retention gains (i.e., more vocabulary will stay in the brain for longer). This could be very beneficial for ecological instructional settings considering teachers can have a longer window of repeating and retrieval. For instance, repetitions might not need to be exactly every 30 days and retrieval might not need to be exactly after 30 days either to obtain the highest retention gains.

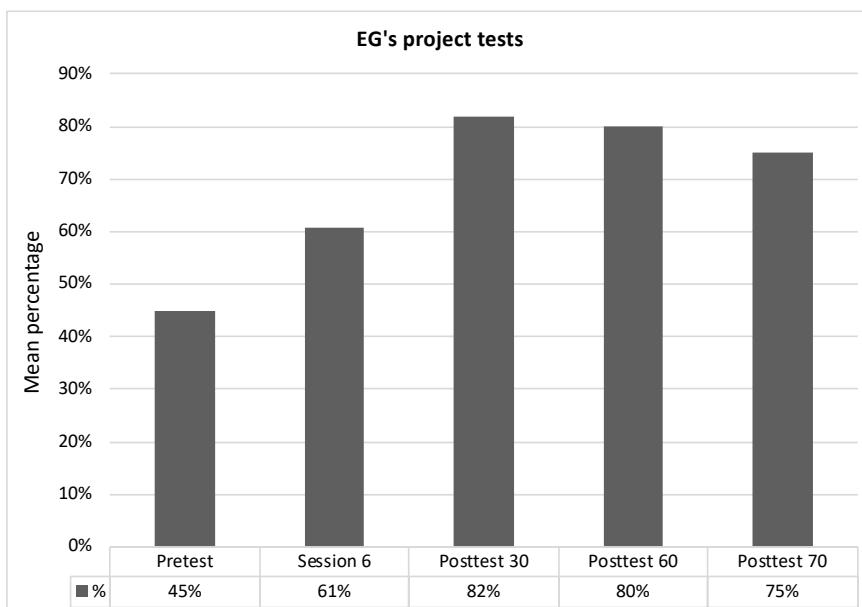
To sum up, considering the actual data resulting from retention tests in this thesis, the highest retention scores were found at 30-day RI given a 30-day ISI. However, by looking at the difference between the retention tests, although more research is needed in this respect, it could still be possible to hypothesize that the optimal RI/ISI combination might not necessarily be an exact value, but rather a range.

#### **6.4 More rehearsals may contribute to better learning and retention**

The notion that more exposition to vocabulary may promote learning has been reflected in the literature (e.g., Nation, 1990; Nation, 2014). Results from the test in session six (see 5.6.2 above) seem to provide evidence suggesting that more vocabulary rehearsals may be conducive to better learning. For instance, in the pre-test there was a general group mean percentage of 45%, while in session six, in contrast, the group mean percentage had already increased to a rounded 61%.

Against previous research (as shown in Figure 6-5 below), results from the test in session six were already among the highest in the graph. This highlights the fact that even after five learning sessions at 30-day ISI and a 90-day RI the experimental group was showing a solid performance increase. From Figure 6-4 it is easy to see the learning progress of the experimental group starting from the pre-test until the 30-day RI test. This suggests an increase in knowledge as participants were going through the learning sessions that is consistent with findings from previous studies. For example, Goossens et al. (2012), conducted a very similar study to Sobel et al. (2011) and

obtained higher retention results. This difference can be partly attributed to the fact that Goossens et al. (2012) employed more learning sessions. In line with this, Fitzpatrick et al. (2008) reported higher scores as the number of learning sessions increased. Consistent with the literature, this research study confirmed those findings in a longitudinal L2 field study with high-school participants. This study found that after eleven learning sessions at a 30-day ISI, participants obtained an average score of 82% in their retention test at 30-day RI. The number of learning sessions appeared to be instrumental in the learning progress and the RI/ISI combination most probably contributed to the fact that even 70 days after the last learning session participants obtained an overall score of 75% (see Figure 6-4 below). These findings are also in line with retention findings from Bahrick (1979). In the same study, the group with three learning sessions at a 30-day ISI obtained an overall average score of 72% when tested at a 30-day RI. The group with six learning sessions (under the exact same conditions) obtained an overall score of 95%, reinforcing the notion that more learning sessions seem to enhance both learning and long-term retention.



**Figure 6-4: EG's project tests**

To conclude, the number of learning sessions seemed to play an important role in vocabulary acquisition and retention. Consistent with previous research, this study found that learning and retention increased as a function of the number of learning sessions.

### **6.5 Vocabulary retention gain**

This thesis has paid special attention to the number of target words that participants were able to retain after learning sessions had stopped. This is of special relevance in language courses where students need to pass an external final exam and their university admissions is at stake. In this situation, retaining vocabulary for long is important, but equally important is the amount of vocabulary they can retain.

The Literature Review exposed the fact that some scholars either concentrated mainly on how long vocabulary could be retained (e.g., Cepeda et al., 2008; Küpper-Tetzel et al., 2014), or on how much vocabulary could be learned in a certain period of time (e.g., Fitzpatrick et al., 2008). Unlike those studies, this project focused on teaching methods for effective and quick vocabulary acquisition, while concentrating also on the amount of and the time that vocabulary was kept in memory.

In order to understand the contribution of the present study to the field, its findings were contrasted against the literature. Table 6-1 below shows (in descending order) the number of target words that participants had to learn through spaced repetition in previous research. The number of learning sessions and the average of new words expected to be learned per session are also included to help with the comparison. The average of target words participants had to learn per learning session helps understand the scope of each one of the studies. All conditions being equal across two studies, for instance, it should in theory be more difficult for participants to learn and retain vocabulary in the study with the larger number of target words. Notice that Cepeda et al. (2008) is also included in the table, but the study actually worked with trivia facts rather than target words.

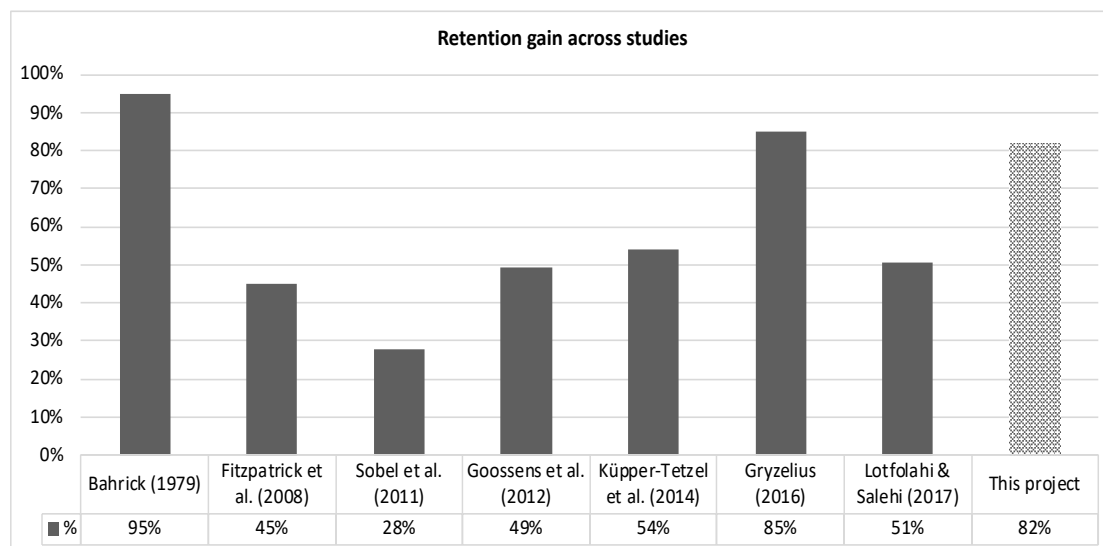
<b>Study</b>	<b>Target words</b>	<b>Learning sessions</b>	<b>Average per session</b>
Fitzpatrick et al. (2008)	300	20	15
Johnson & Heffernan (2006)	112	15	7.46
This study*	100*	11*	9.09*
Bahrck, H. (1979)	50	6	8.33
Cepeda et al. (2008)	32*	2	16
Küpper-Tetzel et al. (2014)	26	2	13
Bloom & Shuell (1981)	20	3	6.66
Gryzelius, (2016)	20	3	6.66
Goossens et al. (2012)	15	4	3.75
Lotfolahi & Salehi (2017)	10	2	5
Sobel et al. (2011)	4	2	2

**Table 6-1: Number of target words (in SR condition) and learning sessions in each study**

The table above shows that in comparison to previous research, this study used one of the highest numbers of target words in the study (100 target words), that participants were expected to learn in eleven learning sessions at an average of 9.09 words per session. Although a high number of words per learning session cannot explain on its own how challenging the acquisition of vocabulary was for participants (e.g., teaching strategies and participant characteristics play a major role here), it can still help comprehend findings better. For instance, Fitzpatrick et al. (2008) used a high number of 300 target words, which does not seem to be excessively large when applied to the context of the study itself in which one subject, a linguistics university professor, was expected to learn 15 new Arabic words per session. In terms of the overall number of target words, the main research project of this thesis could be contrasted more directly against Johnson & Heffernan (2006), but since the authors in that study did not test for retention no comparison can be made. Another study this project could be contrasted directly against is Bahrck, (1979), specially regarding the total number of

words per session. However, special considerations should be made since there were major differences regarding the overall number of target words, quantity of learning sessions and methodology employed (e.g., Bahrick’s was a lab study with young adults as participants). Finally, the other study this project can be more directly contrasted against is Gryzelius (2016), which was the closest in the methodology. Gryzelius (2016) was a field study taking place during regular class hours with 8th grade participants taking Spanish as L2, who learned vocabulary explicitly through the use of an online flashcard program. Although the difference in number of words per session was apparently not so large, Gryzelius (2016), was still a rather short study in duration as the learning period extended only for eight days.

In terms of research findings, Figure 6-5 below shows a summary of retention gain reported by previous research. The table also includes results arising from this study at 30-day RI and provides a graphical representation of how these results compare to other similar spaced repetition studies. Notice that for a more direct comparison, the table includes only studies with retention intervals from 30 to 42 days.



**Figure 6-5: Vocabulary retention gain across studies (30- to 42-day RI)**

The figure above reveals that results from this project are higher than most and even when test scores are not the highest, they still seem to be rather salient considering the other two higher scoring studies cannot be contrasted so straightforwardly. To start with, the highest-scoring study was Bahrick (1979) which extended for six months, at a 30-day ISI and at 30-day RI. Higher results could be expected from this lab study also since, as opposed to field studies, there is a greater control over variables which could facilitate the development of the project and participant learning. Also, even when in the main study of this thesis participants showed high levels of commitment (see 6.6 below), university students (as opposed to younger learners) have been reported to show more motivation than high-school students (e.g., Pintrich & Schunk, 2002; Tüysüz et al., 2010).

The second highest scorer in the figure was Gryzelius (2016), which also reported a very high retention gain. In comparison to the research project of this thesis, Gryzelius (2016) was very short with only three learning sessions, with an interstudy interval varying from one to five days, and with only 20 target words. Despite Gryzelius' (2016) findings, specially at a 39-day RI, it could still be argued that there was less risk of participants forgetting information considering the three learning sessions took place within eight days overall. It would arguably be expected that in a longitudinal field study like the one introduced here, there is more risk of forgetting and participants might lose interest to participate in the study over the long run, which could hinder study results.

When looking back at Table 6-1, of the three studies with 100 or more target words, this study was the one obtaining the highest retention scores. These findings could be attributed to the fact that Fitzpatrick et al. (2008) did not seem to test for retention at the optimal time, and Johnson & Heffernan (2006), never really tested for retention after learning sessions had finished. This project, in contrast, used a dedicated agenda of rehearsals and post-tests aiming at the highest possible retention scores. At the same time, the fact that all target words were introduced and rehearsed since day one, (as opposed to Fitzpatrick et al., 2008) probably also helped with retention. Finally, the use of teaching strategies (e.g., retrievals from memory and writing words down) also seemed to enhance learning.



The comparison above reveals that the main project of this thesis had interesting results considering it was a longitudinal field study that extended for thirteen months overall. At the same time, what also adds to the relevance of this project's findings is that high-school learners seem to be hard to motivate (specially in comparison to university students), and that motivation is hard to sustain for extended periods of time. The other 'longitudinal' field study with young learners as participants in Figure 6-5 above was Küpper-Tetzel et al. (2014). That study, however, had 26 target words, extended only over 20 days, had two learning sessions at 10-day ISI, and obtained a very low 54% retention gain.

To conclude, Figure 6-5 revealed that this project's retention gain was the highest of its kind (a field study extending over thirteen months with high-school students as participants and 100 target words). A blend of an optimal retention interval and interstudy interval combination, together with the use of digital flashcards, retrievals from memory and high subject motivation seemed to reinforce vocabulary acquisition and appeared to have enhanced vocabulary retention. The amalgamation of those strategies seemed to have contributed to retention gains well above previous spaced repetition research studies.

## **6.6 Participant motivation**

Motivation deserves special attention since previous researchers reported it as a major drawback negatively affecting projects' main goals. For example, in Johnson & Heffernan (2006) since vocabulary learned in the project was not part of the course curriculum, the subjects did not seem to be fully interested in the study as they did not see the benefits of learning new words. The exact same situation was also true in the replication study reported in Chapter 3 and in Gryzelius (2016). Lack of participant engagement was also reported in Erbes et al. (2010) in study with high-school learners and in Goossens et al. (2012) with primary-school students.

Although motivating participants was not one of the main goals of this project (i.e., it was not measured in terms of data collection aiming at answering the research questions), different strategies were still envisaged to motivate them, and to avoid lack of interest and engagement. Therefore, two different strategies were employed in

the main project. The first of the strategies was based on findings from Milliner (2013) and it consisted on using (as target words) vocabulary that could help participants be better prepared for their Ab Initio course requirements. The second strategy was based on ideas from Goossens et al. (2012) and it consisted on offering a variety of learning tasks to keep subjects entertained, focused, and actively engaged.

Despite the fact that this project combined strategies from the studies mentioned above, results very different from what those previous researchers had reported. Contrary to those studies, researchers' observations and participants' reflections in this study showed that subjects were positively engaged and highly motivated to participate in the project at all times. Following Milliner's (2013) strategy, the inclusion of target vocabulary that would help participants achieve higher grades in their Ab Initio course seemed to work particularly well as a motivator. Together with that, the constant checking for learning by the researcher and the verbal encouragement to focus and work confidently on the activities seemed to keep participants engaged as well. The idea of offering a variety of activities suggested by Goossens et al. (2012) seemed to work very well for motivation as well. This was probably also the case since participants were allowed to study 'their own way' sometimes, for instance by taking notes or drawing pictures if they felt this helped them memorize words better. Finally, the online interactive activities, especially games with friendly competition against peers helped with motivation almost as much as the chocolate that was sometimes offered as rewards.

Considering the results of the main study in this thesis, it can be argued that the retention gain, as discussed in the previous section, can be attributed to the combination of both the spaced repetition schedule and motivation. Although both those aspects are related, the optimal combination of the retention interval and the interstudy interval still seems to be the most important aspect regarding retention gains. It could be argued that even without motivation, the mere repeated exposure to the words will produce some learning and retention. High levels of motivation alone, on the other hand, could not produce high retention gains at long intervals (e.g., 30-day RI) if there is no appropriate RI/ISI in place as vocabulary will most probably be quickly forgotten.

A research project could help investigate which of the two aspects (the appropriate combination of retention and interstudy intervals or participant motivation) has a major influence in vocabulary retention gains. The study, for example could have three different groups. One group could be taught target vocabulary through a proper combination of intervals (RI/ISI) while keeping motivation high. Another group would be taught the same target vocabulary aiming at high motivation but with a poor RI/ISI combination. And finally, the third group would undergo a treatment with a proper RI/ISI combination but without focusing on participant motivation (for instance through the use of flashcards on paper, rote repetition, and without any interactive activities or games). A comparison of the results of the three groups would help understand the influence that participant motivation and a well-structured spaced repetition strategy may have on long-term vocabulary retention.

To conclude, a good spaced repetition schedule and the motivational aspect to enhance long-term vocabulary retention could be easily implemented in everyday L2 classrooms. For example, in courses such as Spanish Ab Initio, a monthly repetition of a set of target words could most probably reinforce vocabulary retention, specially if learners' motivation is kept high through the use of a variety of interactive activities to engage students, and through the use of activities that replicate those to be found in the final exam.

### **6.7 Limitations of this study and suggestions for further research**

There are a few limitations to this study that need to be considered to comprehend how findings contribute to the spaced repetition field.

To start with, it needs to be considered that target words in the main tests of this study (pre-test, test in session six, and post-tests) appeared in context. The reason for this was that tests could serve two main purposes, i.e., to collect data for the research project, and also as a means to continue to prepare participants to succeed in their compulsory language course requirements (the ability of learners to be able to make meaning from context was a big part of the skills needed to succeed in the Ab Initio course). This, however, could question participant's actual knowledge of the target words. For instance, if participants learn the target words in isolation (together with

their translation into another language for instance) and then are tested in their ability to recognize what the translation is, given a certain word (through a multiple-choice test, for example), it could be very clear to see what words participants know and the ones they do not know. An example of this could be by teaching participants that the Spanish word *hombre* means ‘man’. Later when participants are tested, the word *hombre* appears in the exact same form as it was first presented, and subjects need to select the appropriate translation given four options (e.g., ‘house’, ‘man’, ‘woman’, ‘child’). However, when context is involved, if participants make mistakes in a test, it could be due to the fact that they simply had difficulties with words in context, but not necessarily that they ignored the meaning of the target words in isolation. Therefore, further research could look at testing isolated words to obtain a clearer picture of participant vocabulary acquisition and retention. Although this would eliminate the resemblance of the activities in the study to what learners would find in the final exam (specially in the case of a Spanish Ab Initio course), this would help obtain more precise data. This would provide a better picture regarding the actual target words participants knew before the treatment, and the ones participants can retain after the rehearsals without worrying about the extra variables (e.g., verbal conjugation and number and gender agreement) added by words appearing in context. It is still important to bear in mind, however, that the fact that words appeared in context in tests in this project helped participants see the connection between the study and their actual course curriculum. This seemed to contribute enormously to motivation and engagement during the study (which as discussed in the previous section can negatively affect study results). As a consequence, future research should also consider a different method to keep subjects motivated such as adding more activities in the project where participants can practice the target words and also practice skills needed to pass course requirements (e.g., further reading comprehensions with questions that follow the same format as the course final exam).

A second issue that limits the effect of these findings on general language courses is the fact that participants were students at an international school which can be deemed as very particular in several ways. Students appeared to be rather pampered and were offered lots of opportunities to succeed in every class. For instance (at least at the school where this project took place), there was, by rule, a limited number of students

per class (maximum of fifteen), who had access to state-of-the-art devices (the latest technology in audio, visual and computer devices were constantly used for language learning) and were offered full support for learning (e.g., after class support, homework club, counselors). This represents the reality of a small population; therefore, future research should consider replicating this study with a different sample (e.g., in public schools or private language schools). This would ideally provide a clearer picture of whether findings from this study expand also to a much larger population of language learners. For instance, this study could be expanded by implementing spaced repetition into an ongoing language course in a public high-school. Several adjustments would need to be made considering the multiple differences to be found in such a context. For instance, course requirements, number of course hours per week and course expectations would probably be very different to the Spanish Ab Initio course referred to in this thesis. Therefore, the replication study might need to consider a different number of learning sessions, the target words to be used and how many of them. A topic that would need special attention in a replication study is the number of participants, behavioral issues, and learners' motivation and engagement. In a public educational system (as opposed to international schools), with a larger population of students per class, it would most probably be necessary to plan a very strong motivational strategy to keep participants engaged at all times, specially in a longitudinal study.

Another important limitation of this research project lies in the fact that the control group was tested on vocabulary they had probably never seen or, if they had, they had most probably come across it only incidentally. Therefore, a way to avoid this situation between the experimental group and the control group, target words could be taught to both groups. The main difference would be that the experimental group would follow the spaced repetition condition, while the control group would learn the target words together with any other word they learned during the regular course of studies. For instance, during their regular Spanish Ab Initio course, the control group would follow the curriculum, as always, covering topics such as City life, Country life, Health, Food, Holidays, etc. (see 4.4 above). Typical topic words (about 75 per unit) would be introduced in a dedicated list (for an example see Appendix VII) and through reading activities with a final unit revision and exam. A typical lesson would

include reading and writing activities, videos, and the use of Quizlet with also some grammar drills. The replication study could follow a double-blind method in which the control group and the teacher leading the group would never know what the final study target words are. For instance, the researcher could prepare a vocabulary list (similar to the one found in Appendix VII) for each unit the control group would cover during the duration of the treatment. The researcher could add about 20 new words to each list including some study's target words (related to the unit topic) plus some distractors. This way the collaborator instructor and subjects in the control group would be exposed to the study's target words, but without knowing exactly which ones they are so no extra attention could be paid to them. This would definitely require a collaborator teacher willing to help without interfering with the project, for instance by following instructions, and by not trying to compete against the experimental group (i.e., without changing teaching habits so the control group would score higher in tests). This way both groups would be taught the target words at one point and the retention results arising from replication post-tests would be more precise. Considering the type of repeating used in Spanish Ab Initio courses (see 1.3 above), participants in the control group would have several expositions to the words. It can only be speculated, however, that considering a spaced repetition study such as the one introduced in this thesis, the experimental group would still most probably have higher retention gains than the control group.

When it comes to retention intervals and interstudy intervals there are also certain limitations of this study that need to be addressed. For example, another issue that can also question the validity of findings from this study is the length of the interstudy intervals (ISI). This can be rather controversial especially when discussing whether interstudy intervals (ISI) should be equal in length to the retention interval (RI), or whether the interstudy interval should be only a portion of retention interval. Although the interstudy intervals were planned to be 30 days in length in this project, due to school intricacies many of them had to be modified to accommodate them to the school calendar. The average interstudy interval of all eleven learning sessions, therefore, resulted in 34.7 days. Thus, further research in this field could contribute by providing more precise data. For example, a similar ecological study could be conducted extending over a whole academic year. The study could start at the

beginning of the academic year, and finish at the end. This study could still be integrated into an ongoing language course and would be ideal to keep a constant interstudy interval (every 30 days). Although this however would lack the actual dynamics of a two-year course (including the summer break), this would still provide relevant data in a longitudinal field study extending for about ten months. This would help to establish a greater degree of accuracy on this matter by providing more precise data regarding the optimal RI/ISI combination with learning sessions taking place at the exact same interval every time.

Despite the fact that findings from this thesis suggest that given a 30-day RI, it is best to repeat every 30 days, there is abundant room for further research in determining how exactly vocabulary is retained. Further data could be collected in this matter by adding some other retention intervals. For instance, a 0-day RI post-test could provide important data regarding how forgetting starts to set in after rehearsals have stopped. This is a controversial issue considering in Bahrick (1979), after six learning sessions at 30-day ISI, a 30-day RI post-test showed an increase in test results comparing the test immediately following the last learning session (at 0-day RI) and the post-test 30 days after. This showed that participants were scoring higher after rehearsals had stopped. This case was described as the inverted-U of retention (Bahrick, 1979 and Cepeda et al., 2006) where the optimal RI/ISI combination produces the highest retention scores, but retention would be lower before and after that. A very different situation was found in this study and in Fitzpatrick et al. (2008) where test scores showed a quick decrease in test results between the last learning session and the first post-test. Therefore, further research could expand this study by adding two other post-tests to its testing schedule. This could provide data regarding how much vocabulary participants have learned in the treatment and can remember at 0-day RI. The first test should occur immediately after the last learning session, and another post-test should occur at 15-day RI. This could also provide relevant data regarding two different facts. For instance, it could provide relevant data regarding the retention curve, whether after rehearsals test scores are higher or lower than a 0-day RI test. The second issue this could also help elucidate, is the fact that, as it was exposed in 6.3 above, for a certain interstudy interval there might not be an exact time at which

retention is at its highest, (e.g., a 30-day ISI might produce similar results at 30-day RI, but this could also extend a few days before and after that).

Another issue exposed in this thesis regarding the combination between the retention interval and the interstudy interval (RI/ISI) is the fact that the interstudy interval might not necessarily need to be always similar in length to the retention interval. Information does not seem to be retained for long when interstudy intervals are too short, but if the intervals are too wide apart information appears to be forgotten between rehearsals. According to findings from this thesis the retention interval and the interstudy interval should be equal in length for vocabulary to be kept in memory. This seems to be appropriate for the 30-day RI and 30-day ISI of this thesis. However, if the retention interval is 365 days in length or more, does it mean that interstudy intervals should also be 365 days in length or more? This should also be explored further to discover how wide apart rehearsals could still be without running the risk of forgetting information altogether. Therefore, further research should concentrate on investigating different lengths of interstudy intervals to determine what the maximum possible lapse could be between rehearsals before information is (mostly) forgotten and needs to be learned completely again. Interestingly enough, if information is indeed forgotten between interstudy intervals that are too wide apart, then the interstudy interval might need to be a portion of the retention interval, which then would be in line with findings from Cepeda et al. (2008) and Küpper-Tetzl et al. (2014). If this is the case then, the optimal combination between retention interval and interstudy interval might need to be explored even further. At this point, it can only be speculated that the retention interval and the interstudy interval might need to be equal in length when the retention interval is not too large (30 days in length). Otherwise, in cases when the retention interval is longer than 30 days, the interstudy interval might need to be only a portion of the retention interval to avoid forgetting between rehearsals.

A final topic that future research should consider is the fact that this thesis focused on how much as well as how long vocabulary can be retained. This is a polemical issue that should be carefully considered specially when investigating the actual implementation of spaced repetition in educational environments. Studies focusing on



how much vocabulary is retained, should pay special attention to the tests themselves. For instance, if most participants score 100%, the test was probably too easy, not showing exactly what the scores could have been if there had been a higher ceiling. Therefore, researchers should avoid a ceiling effect by making tests that allow participants show their potential without being restrictive. On the other hand, as opposed to the researcher, the language teacher will be more interested in the practical application of spaced repetition into everyday classrooms. It could be generalized that language teachers look at methods to improve general learning. However, teachers in most educational institutions, instead, seem to focus almost exclusively on test results, since teaching capabilities are measured by learners' grades in external exams (Fournier-Kowaleski, 2005). With this in mind researchers should consider tests with a high ceiling that could show participants full potential in a study. For example, a study could be conducted in which a very large number of target words are used (specially in comparison to what learners would typically learn in an ordinary language course). This could help the researcher avoid a ceiling effect, and at the same time, the investigation could provide relevant data regarding retention with a very high number of words. This data could later be applied to actual language courses where the number of target words is usually less, so learners could manage to retain the maximum number of target words leading to high grades in tests. This could be very attractive for the language teaching field since from a practical standpoint it could help improve students' scores in tests.

## **6.8 Conclusion**

This chapter has interpreted project's results in order to address the most salient topics in this study. Findings suggest that the methodology employed in this project fostered vocabulary acquisition and strengthened long-term retention.

To start with, the fact that the experimental group scored significantly higher than the other groups revealed that findings are not only significant against a very similar group (the control group), but also that they can extend to a wider sample (as it was the case of the historical group). This evidenced the fact that Spanish Ab Initio students, for example, could benefit greatly from the use of spaced repetition by improving vocabulary acquisition and vocabulary retention at the end of their two-

year course. For instance, in general in this type of course, where about 1200 new words are learned over the two-year period at a rate of six words per contact hour, the overall course word gain could be initially expanded to 1500 words. The first month of the course could be dedicated to learning all 1500 words explicitly through the use of bilingual digital flashcards. Starting from the second month, the typical course (with readings, speaking activities, grammar drills, etc.) would begin with one or two topics being introduced per month. Three consecutive lessons per month could be exclusively devoted to an explicit revision of all target words. At the same time, curriculum designers should create a series of readings where all of the target words are included aiming at implicit exposure of target words. Those readings could be worked in class at a rate of three readings per month. This combination of explicit teaching plus constant incidental exposure to target words would most probably contribute to support learning and to retain large amounts of vocabulary until the final exam. The initial 1500 target words could be expanded to 2000 in future courses if learners respond well to the initiative.

Second, the third research question was controversial as it investigated the best possible RI/ISI combination. Findings suggest that the interstudy session appears to provide higher retention scores when it is equal in length to the retention interval. What was even more revealing was the fact that retention seemed to be equally high before and immediately after the 30-day RI. This suggested that perhaps repeating vocabulary every 30 days can provide similar retention gains in a period extending from about 20 to 40 days after the last learning session. This would be in line with the inverted-U shape theory of retention referred to in Bahrick (1979) and Cepeda et al. (2006). The theory stated that for a given RI/ISI combination, the optimal combination of intervals produces the highest retention peak, but scores are lower before and after it.

The third issue arising from this chapter, which actually further explains the argument stated in the previous paragraph highlights the importance of future research to continue to investigate the optimal combination between the retention interval and the interstudy interval. Results from the test in session six in this study brought to focus an inverted-V curve of learning and retention. This suggested that through the

learning process participants score higher as the treatment progresses. The highest peak of knowledge is acquired at the very last learning session, but retention begins to decline after that (as reported by this study and Fitzpatrick et al., 2008). The fact that some vocabulary will still be lost after rehearsals have finished seems to be inevitable and the actual curve that retention will follow is still to be elucidated.

A fourth salient issue was that (as reported by the literature, the replication study and findings from this study) the number of learning sessions might be another important factor that can enhance long-term vocabulary retention. More rehearsals should contribute to better acquisition and higher retention levels.

Finally, as stated above (see 6.6) although participant motivation was not part of the research questions in this study, it still deserved special attention considering it could negatively affect a project's results. The teaching strategies and materials used in this thesis (e.g., online flashcard platforms, interactive activities, games, variation of the activities for learning) appeared to help keep participants motivated and engaged at all times. All of this seemed also to have contributed to better learning, and incidentally to better retention.

To sum up, this chapter has made evident the fact that findings from this project can make a contribution to the spaced repetition field. At the same time, the limitations exposed still suggest that further research should be undertaken to continue to comprehend how spaced repetition can enhance large amounts of vocabulary to be retained for longer. In terms of implementing spaced repetition into the everyday classroom, curriculum designers and teachers could improve vocabulary learning and retention by making a few adjustments to current teaching methods. In courses such as Spanish Ab Initio (which rely heavily on vocabulary knowledge) when planning the lessons before the course begins, the teacher should first start by looking at the time when learners will sit the final exam. This day and the day of the last learning session are crucial to determine the retention interval (which in general tends to be from 15 to 50 days). The next step would be to plan the interstudy sessions, which according to findings from this thesis should ideally be equal in length to the retention interval. This in general translates into repeating the target words about once a month.

It is of particular importance to plan the re-learning sessions considering the context and schedule of the academic institution (e.g., breaks, field trips, celebrations) so repetitions can be adjusted accordingly. The next step is to ensure that students acquire vocabulary properly. Explicit learning methods (through the use of online flashcard platforms, for instance) appear to be very efficient in this matter and should be used to introduce and ensure vocabulary is encoded properly. Later, the use of implicit methods (such as readings and videos) should be used to increase exposure that will eventually enhance retention. Finally, the use of interactive activities and games help keep students motivated, which is key to ensure they do their best and learning and retention take place as expected.

To conclude, this thesis investigated the efficacy of spaced repetition to improve vocabulary learning and enhance long-term retention. An important issue that arose from this thesis is the fact through a series of repeated exposures vocabulary can be learned successfully. Although learning is crucial for retention, on its own, proper learning cannot guarantee that vocabulary will be retained for long. The proper arrangement and combination of retention and interstudy intervals promote retention. Finally, the use of online flashcards seemed to enhance vocabulary acquisition and boosted participant motivation through interactive activities and games. Participant motivation in spaced repetition research studies emerged as crucial factor that can negatively influence study results.

## Chapter 7: Conclusion

This thesis investigated the efficacy of spaced repetition to enhance long-term vocabulary retention. The first question guiding this thesis was how vocabulary retention gains through spaced repetition compared against those of traditional teaching methods. This thesis also examined whether a 30-day interstudy interval (ISI) would produce higher vocabulary retention gains at a 30-, 60- or 70-day retention interval (RI).

The focus of attention of this thesis was directed towards vocabulary retention considering two major reasons. To start with, vocabulary is one of the most important components in foreign language learning considering large amounts of it are needed for effective communication (e.g., Wilkins, 1972; Meara, 1980; Hu and Nation, 2000; Schmitt, 2019), and the main difference between language learners and native speakers seems to be, precisely, the amount of vocabulary known (Laufer, 1998). The second reason was the fact that some language courses (such as IB Spanish Ab Initio) focus mainly on vocabulary knowledge, hence it is crucial to acquire a relatively large number of vocabulary items in a short period of time. Teaching large amounts of vocabulary in such a course could be a straightforward solution. However, as stated in Bahrick (1979), most of the information we learn is forgotten. This exposed the need to employ teaching strategies that would ensure both, that large amounts of vocabulary are learned, and that vocabulary would stay in memory for long. The research community have addressed this issue and found that repeating vocabulary enhances learning and avoids forgetting. If that vocabulary is meant to be retained for extensive periods of time, then rehearsals should be spaced in time (rather than being massed without lags in between). However, despite extensive evidence suggesting the fact that spaced repetition enhances vocabulary retention, it has not been implemented widely in language courses yet (Dempster, 1988; Kornell, 2009; Sobel et al., 2011; Schmitt, 2019).

One of the main reasons for this lack of implementation seems to be the absence of consensus regarding when exactly to repeat. To start with, researchers agree that when spacing repetitions there are two main lags that need to be determined. Those

lags are the interstudy interval (ISI), which is the lag between vocabulary rehearsals, and the retention interval (RI), which is the time lag between the last rehearsal and the moment when information will be retrieved from memory. This however, (see 2.1.6 above) generated a major controversy regarding the length of the interstudy interval and the retention interval. Scholars do seem to agree on the fact that the interstudy interval (ISI) increases as a function of the retention interval (RI), i.e., the longer the retention interval, the longer the interstudy interval. However, there seems to be a major disagreement in the literature regarding the exact combination between the interstudy interval and the retention interval. While some researchers suggest that the retention interval and the interstudy interval should be equal in length for higher retention gains, others state that the interstudy interval is always a portion of the retention interval. That portion of the retention interval is something scholars still do not fully agree upon either.

The second reason for the lack of implementation of spaced repetition in real classrooms seems to be that most spaced repetition studies in the past were conducted in labs, were short in duration and/or had university students as participants. Therefore, more longitudinal research is needed investigating how spaced repetition applies to young learners in more ecological environments.

This thesis, then, was conceived with the aim of filling those gaps and it was original since it combined six different aspects. First, it introduced a spaced repetition field study that extended over 13 months (considering only the learning period) having high-school learners as participants. Second, target vocabulary was part of the suggested vocabulary participants needed to learn in order to succeed in the IB Spanish Ab Initio course. Third, this project employed innovative teaching methods (through the use of online flashcard platforms with a variety of interactive activities and games), while also trying to keep the environment as ecological as possible. Fourth, retention test scores of participants being taught through spaced repetition (SR) were contrasted against other groups being taught the traditional way (the way the course had been traditionally taught at the school where the project was conducted), rather than contrasting it against massed repetition as seen in previous research (e.g., Bloom & Shuell, 1981; Sobel et al., 2011; Gryzelius, 2016; Lotfolahi

& Salehi, 2017). Fifth, the project also aimed at contributing to the field by providing further data regarding the optimal RI/ISI combination and retention curve. Finally, this project also focused not only on how long vocabulary was retained, but also on how much vocabulary was retained at different retention intervals.

### **7.1 Findings regarding SR and RI and ISI lags**

Starting with the replication study (see Chapter 3: above), it can be clearly seen that repetition alone cannot guarantee long-term vocabulary retention. It is the previous planning and the allocation of the interstudy intervals (ISI) according to a certain retention interval (RI) that should guide the learning session schedule. Although both Johnson & Heffernan (2006) and my replication of it failed to provide conclusive data regarding how much information was kept in memory (after rehearsals had stopped), the replication study seemed to offer a hint. An analysis of all three tests in the study still showed some retention tendency. Figure 3-8, for instance, revealed that the results from the retention test were the lowest in the study. This is an unexpected finding considering that Table 2-13 above showed that all of the studies analyzed in the Literature Review reported higher retention scores in comparison to pre-test scores. This suggests that apart from participant lack of interest to do well in the test, the low retention scores could also be attributed to the fact that the RI/ISI combination had not been properly implemented.

Searching for the optimal combination between the interstudy interval and the retention interval (RI/ISI), I investigated which retention interval (RI) would provide the highest results given an interstudy interval (ISI) of 30 days. As discussed in 6.3 above, results from this study are in line with Bahrick (1979) and Lotfolahi & Salehi (2017) that claim that the interstudy interval and the retention interval should be equal in length for best retention gains. What may also be interpreted from the same findings is that for a certain interstudy interval, the optimal retention interval might not necessarily be an exact number, but rather an approximation. For instance, provided a 30-day ISI, highest retention scores might be obtained at 30-day RI as well as a few days before and after that. At the same time however, findings from this thesis seemed to expose two other issues that might need further research in order to provide more conclusive evidence. There appears to be an inverted-V shape of

learning and retention arising from an analysis of the results of all tests in the project. This would mean that during the learning process, as vocabulary is acquired, vocabulary knowledge will increment gradually reaching its peak at the last learning session. From that moment on, during the retention period (after rehearsal have stopped), vocabulary knowledge decreases. The line tracing the forgetting of vocabulary will drop sharply or more slowly, depending on the interstudy interval that was used during the learning sessions. This concept, deserving more research, challenges findings from Bahrick (1979) where retention results (at 30-day ISI and 30-day RI) were higher than results from the last learning session. In that study then the retention line continued to rise after the last learning session, rather than dropping in the retention period. However, Bahrick (1979) also reported that the 0-day ISI and the 1-day ISI conditions reported findings that are in line with this thesis, suggesting, therefore, that more research is needed to provide more conclusive evidence. The final issue that also arose in this thesis is the fact that the number of learning sessions might also play an important part in long-term vocabulary retention. Although more learning sessions could ideally enhance learning and (if applied properly) longer retention, the exact number needed to affect retention results might still need to be determined. For instance, it would be interesting to see whether (when aiming at long-term vocabulary retention) there is a correlation between number of learning sessions, the length of the retention interval and the interstudy session intervals. To conclude, although spaced repetition, the retention interval, the interstudy interval and the number of learning sessions do play an important part in long-term vocabulary retention, it seems that a general formula (if it exists) is still to be determined. It could be speculated that considering evidence from previous research and findings from this thesis, the number of learning sessions do play an important role (e.g., Bahrick, 1979; Fitzpatrick et al., 2008; Nakata, 2017).

## **7.2 Project's findings regarding teaching and learning**

It would be wrong to say that vocabulary retention depends solely on the appropriate combination of the retention interval and the interstudy interval as there would be a major component missing. Based on the notion that if information is not learned properly, it will not be retained properly (e.g., Bahrick, 1979; Nation, 1990), what is



left then is the discussion of the actual teaching that also plays a decisive role in how much and how long that vocabulary would remain in the brain.

The amount of vocabulary retained is equally, if not more important than the time that vocabulary is retained for. In courses such as Spanish Ab Initio obtaining a high grade is crucial for university admissions. Therefore, special attention should be taken when deciding the strategies used to teach large amounts of vocabulary in a rather short period of time, specially when no homework can be assigned so learners can be exposed to the target language also outside of the classroom. For instance, in this type of language course in general students are expected to learn an annual average of around 600 words, at a rate of six words per contact hour. Hence, preparing complete L2 beginners to comprehend authentic materials in a two-year language course seems to be a daunting enterprise. This seems specially complex also considering previous research (Nation, 2006) has reported that about 8000 to 9000 word families are needed to comprehend authentic texts.

This thesis has shown that high retention gains can be achieved through a combination of teaching strategies (specially in comparison to similar spaced repetition studies as seen in Figure 6-5 above). To start with, the use of online flashcard platforms seems to be very effective not only for teaching per se, but also to motivate and engage learners through interaction and variation of catchy activities and games. The use of vocabulary that could help participants obtain high grades as target vocabulary in the study, also seemed to be a great motivator for subjects. At the same time, the use of authentic materials (as seen in the replication study), not only to prepare learners for their tests, but also as a means to connect the course to the real world, appeared to keep participants interested in learning (especially if authentic materials are movie trailers, which on their own seemed to succeed in attracting learners' attention). The constant monitoring to check that participants were on track and actually acquiring the required vocabulary, also proved to be particularly effective as well. Finally, as seen in the literature and in both studies in this thesis, motivation plays a crucial role to engage students and ensure that learning takes place successfully.

### **7.3 Pedagogical implications**

Although findings from this thesis contributed to the claim that spaced repetition can enhance long-term vocabulary retention and more research is still needed in this matter, it could still be claimed that spaced repetition could be implemented in language courses without much adjustment to everyday teaching.

It seems that in courses where there is a monthly repetition in place, for instance, distributing vocabulary rehearsals over time could be easy to implement. Although the optimal combination between the retention interval and the interstudy interval is yet to be determined, (as discussed in 6.3 above) research has shown that given a retention interval of a certain length, and then repeating at the same length can contribute to retain vocabulary for longer. Therefore, as it was seen in this project, repeating every 30 days seemed to contribute greatly to the retention of vocabulary 30 days after the last rehearsal in which on average, 82% of the target words were remembered (if tests results are carefully extrapolated, that is). At the same time, it seems that blending everyday lessons with explicit vocabulary teaching could also bring great benefits to both learning and retention. Acquisition could surely be improved, especially if online flashcard platforms are employed taking advantage also of the variety of interactive activities and games the websites offer.

In language courses where students need to remember vocabulary (they have learned before) when they are no longer attending classes, then two major courses of actions should be considered (see 6.8). The first one is to decide when students are going to need to use the words learned, and then lessons in which vocabulary is repeated should be planned accordingly. For instance, if students are taking a final exam at the end of the course, then the time between the last lesson and the exam itself should be considered. If that time is thirty days, for instance, then there should be backward planning so that vocabulary is repeated every thirty days for a series of lessons. The second course of action is to implement teaching strategies that ensure efficient vocabulary acquisition, since if vocabulary is not learned properly, it will not be remembered. For instance, an online flashcard platform could be employed for quick and efficient learning and for engagement to help keep students motivated.

Finally, despite the fact that researchers still need to continue to investigate how to best implement spaced repetition, it is the entire educational community that needs to take a step to finally see spaced repetition systematically integrated in everyday language courses. As stated by Schmitt (2019) curriculum designers and textbook writers should also inspect how to best implement spaced repetition so language teachers also become aware of its benefits and potential implementation in the classroom.

#### **7.4 Final conclusion**

This thesis examined the efficacy of spaced repetition to enhance long-term vocabulary retention in a longitudinal field study with high-school language learners. Findings from this thesis are in line with previous spaced repetition research stating that spaced rehearsals contribute to retain vocabulary for longer.

Two main issues arose as crucial to enhance retention. The first issue is the need for appropriate interval planning. The optimal combination of retention interval and interstudy interval is needed to heighten the retention of large amounts of vocabulary for extended periods of time. This is of major relevance since even in cases of efficient learning, if the optimal combination of the interstudy and the retention interval was not determined properly, vocabulary might still not remain in memory for long.

The second issue that appeared to either enhance or diminish long-term vocabulary retention is efficient teaching strategies. This highlighted the fact that if vocabulary is not acquired properly retention will be negatively affected.

Hence, this thesis has contributed to the field by introducing a longitudinal spaced repetition field study with high-school learners. Findings revealed the importance of the retention interval as a guideline to determine the frequency of the interstudy intervals. At least in the context of this study, it seemed best to repeat every thirty days to guarantee high retention gains thirty days after the last rehearsal. At the same time, motivation arose as a major component that could be detrimental for learning and retention. Finally, this thesis referred to learning and retention as two different

periods. During the learning process rehearsals enhance encoding. The retention period is the time when forgetting starts to set in after rehearsals have stopped.

To conclude, considering there are still several issues in need of further clarification, further research should continue to investigate how long-term vocabulary retention can be further enhanced by spaced repetition. Finally, considering findings from the literature and the fact that spaced repetition could be seamlessly integrated into everyday language classrooms, it is possible to envisage that spaced repetition may change from a research interest to a more practical application implemented in ecological environments.

## Bibliography

- Allum, P. (2004). Evaluation of CALL: Initial vocabulary learning. *RECALL*, 16(2), 488–501.
- Anderson, J., & Jordan, A. (1928). Learning and retention of Latin words and phrases. *Journal of Educational Psychology*, 19, 485-496.
- Ary, D., Jacobs, L., Irvine, C., & Walker, D. (2018). *Introduction to research in education*. Cengage Learning.
- Bahrick, H. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General*, 108, 296-308.
- Bahrick, H. (2005). The Long-Term Neglect of Long-Term Memory: Reasons and Remedies. (A. Healy, Ed.) *Decade of behavior. Experimental cognitive psychology and its applications*, pp. 89-100.
- Bahrick, H., & Hall, L. (1991). Lifetime maintenance of high school mathematics content. *Journal of experimental psychology: General*, 120(1), 20.
- Bahrick, H., & Hall, L. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language*, 52, 566–577. doi:10.1016/j.jml.2005.01.012
- Bahrick, H., & Phelps, E. (1987). Retention of Spanish Vocabulary Over 8 Years. *Journal of Experimental Psychology: Learning, memory, and cognition*, 13(2), 344-349.
- Bahrick, H., Bahrick, L., Bahrick, A., & Bahrick, P. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science*, 4(5), 316-321.
- Bjork, R. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.). *Attention and performance XVII: Cognitive*

*regulation of performance: Interaction of theory and application*, 435–459.  
Cambridge, MA: MIT Press.

- Bloom, K., & Shuell, T. (1981). Effects of Massed and Distributed Practice on the Learning and Retention of Second-Language Vocabulary. *The Journal of Educational Research*, 245-248.
- Bogaards, P., & Laufer, B. (2004). *Vocabulary in a second language: Selection, acquisition, and testing* (Vol. 10). John Benjamins Publishing.
- Borman, G., & Boulay, M. (2004). *Summer learning: Research, policies, and programs*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
- Brodie, R. (2009). *Virus of the Mind*. Hay House, Inc.
- Bunce, D., Flens, E., & Neiles, K. (2010). How long can students pay attention in class? A study of student attention decline using clickers. *Journal of Chemical Education*, 87(12), 1438-1443.
- Caine, G., & Caine, R. (1991). *Making connections: Teaching and the human brain*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Carpenter, S., Cepeda, N., Rohrer, D., Kang, S., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educational Psychology Review*, 24(3), 369-378.
- Cenoz, J., Hufeisen, B., & Jessner, U. (2003). *The multilingual lexicon*. Dordrecht: Kluwer Academic Publishers.
- Cepeda, N., Coburn, N., Rohrer, D., Wixted, J., Mozer, M., & Pashler, H. (2009). Optimizing distributed practice: theoretical analysis and practical implications. *Experimental Psychology*, 56, 236–246.
- Cepeda, N., Pashler, H., Vul, E., Wixted, J., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological bulletin*, 132(3).

- Cepeda, N., Vul, E., Rohrer, D., Wixted, J., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science, 19*, 1095–1102.
- Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research, 66*(3), 227-268.
- Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*(34), 213–238.
- Dawkins, R. (2006). *The selfish gene: 30th anniversary edition*. Oxford: Oxford University Press.
- De Groot, A., & Van Hell, J. (2005). The learning of foreign language vocabulary. *Handbook of bilingualism: Psycholinguistic approaches*, 9-29.
- Dellarosa, D., & Bourne, L. (1985). Surface form and the spacing effect. *Memory & Cognition, 12*(6), 529-537.
- Dempster, F. (1988). The spacing effect: A case study in the failure to apply the results of psychological research. *American Psychologist, 43*, 627–634.
- Dempster, F. (1989). Spacing effects and their implications for theory and practice. . *Educational Psychology Review, 1*(4), 309-330.
- Dietrich, R. (1989). Nouns and verbs in the learner's lexicon. In H. W. Dechert (Ed.). *Current trends in European second language acquisition research*, 13-22.
- Donovan, J., & Radosevich, D. (1999). A meta-analytic review of the distribution of practice effect: now you see it, now you don't. *Journal of Applied Psychology, 84*, 795–805.
- Ebbinghaus, H. (2013). Memory: a contribution to experimental psychology. *Annals of neurosciences, 20*(4), 155-6.
- Elgort, I. (2007). *The role of intentional decontextualised learning in second language vocabulary acquisition: Evidence from primed lexical decision tasks*

*with advanced bilinguals*. Wellington, New Zealand: (Unpublished doctoral dissertation). Victoria University of Wellington.

- Elgort, I. (2011). Deliberate learning and vocabulary acquisition in a second language. *Language Learning, 61*(2), 367–413.
- Ellis, N. (1995). The psychology of foreign language vocabulary acquisition: Implications for CALL. *Computer Assisted Language Learning, 8*(2-3), 103-128.
- Ellis, N., & Larsen-Freeman, D. (2006). Language emergence: Implications for applied linguistics. *Applied Linguistics, 27*(4), 558–589.
- Erbes, S., Folkerts, M., Gergis, C., Pederson, S., & Stivers, H. (2010). Understanding how cognitive psychology can inform and improve Spanish vocabulary acquisition in high school classrooms. *Journal of Instructional Psychology, 37*(2), 120-133.
- Fitzpatrick, T., Al-Qarni, I., & Meara, P. (2008). Intensive vocabulary learning: A case study. *Language Learning Journal, 36*(2), 239–248.  
doi:10.1080/09571730802390759
- Fournier-Kowaleski, L. (2005). *Depicting washback in the intermediate Spanish language classroom: A descriptive study of teacher's instructional behaviors as they relate to tests*. (Doctoral dissertation) The Ohio State University.
- Glenberg, A. (1976). Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. *Journal of Verbal Learning and Verbal Behavior, 15*, 1-16.
- Glenberg, A. (1979). Component-levels theory of the effects of spacing of repetition on recall and recognition. *Memory & Cognition, 7*, 95–112.
- Goossens, N., Camp, G., Verkoeijen, P., Tabbers, H., & Zwaan, R. (2012). Spreading the words: A spacing effect in vocabulary learning. *Journal of Cognitive Psychology, 24*(8), 965-971.



- Griffin, G. & Harley, T. (1996). List learning of second language vocabulary. *Applied Psycholinguistics*, 17, 443-460.
- Gryzelius, T. (2016). El aprendizaje distribuido como estrategia didáctica en la enseñanza del vocabulario de ELE: Un acercamiento a su uso en el salón escolar sueco.
- Hall, C. (2002). The automatic cognate form assumption: Evidence for the parasitic model of vocabulary development. *IRAL*, 40, 69–87.
- Hattie, J., & Yates, G. (2014). *Visible Learning and the Science of How We Learn*. London: Routledge.
- Hill, M., & Laufer, B. (2003). Type of task, time-on-task and electronic dictionaries in incidental vocabulary acquisition. *IRAL*, 41(2), 87–106.
- Hintzman, D. (1976). Repetition and memory. In: Bower, G. (Ed.). *The Psychology of Learning and Memory*, 47-91.
- Ho, P. (2009). *Dynamics of long-term forgetting*. Stanford University.
- Horst, M. (2005). Learning L2 vocabulary through extensive reading: a measurement study. *Canadian Modern Language Review*, 61(3), 355-382.
- Horst, M., & Collins, L. (2006). From Faible to Strong: How Does Their Vocabulary Grow? *The Canadian Modern Language Review/La Revue canadienne des langues vivantes*, 63(1), 83–106.
- Horst, M., & Meara, P. (1999). Test of a model for predicting second language lexical growth through reading. *The Canadian Modern Language Review/La Revue canadienne des langues vivantes*, 56(2), 308-328.
- Horst, M., Cobb, T., & Meara, P. (1998). Beyond a Clockwork Orange: acquiring second language vocabulary through reading. *Reading in a Foreign Language*, 11(2), 207-223.

- Houston, J. (1966). List differentiation and distributed practice. *Journal of Experimental Psychology*, 72, 477-478.
- Houston, J., & Reynolds, J. (1965). First-list retention as a function of list differentiation and second-list massed and distributed practice. *Journal of Experimental Psychology*, 69, 378-392.
- Hsueh-Chao, M., & Nation, I. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13(1), 403–430.
- Hu, M., & Nation, I. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language*, 23(1), 403–430.
- Hunt, A., & Beglar, D. (2005). A framework for developing EFL reading vocabulary. *Reading in a foreign language*, 17(1), 23-59.
- IB. (2002). Spanish Ab Initio - Language Specific Syllabus. International Baccalaureate Organization.
- Jacoby, L. (1978). On interpreting the effects of repetition: solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior*, 17, 649–667.
- Jiang, N. (2002). Form-meaning mapping in vocabulary acquisition in a second language. *Studies in Second Language Acquisition*, 24, 617–637.
- Joe, A. (1998). What effects do text-based tasks promoting generation have on incidental vocabulary acquisition? *Applied Linguistics*, 19(3), 357–377.  
doi:10.1093/applin/19.3.357
- Johnson, A., & Heffernan, N. (2006). The Short Readings Project: A CALL reading activity utilizing vocabulary recycling. *Computer Assisted Language Learning*, 19(01), 63-77.

- Küpper-Tetzel, C., & Erdfelder, E. (2012). Encoding, maintenance, and retrieval processes in the lag effect: A multinomial processing tree analysis. *Memory, 20*, 37-47.
- Küpper-Tetzel, C., Erdfelder, E., & Dickhäuser, O. (2014). The lag effect in secondary school classrooms: Enhancing students' memory for vocabulary. *Instructional Science, 42*, 373-388. doi:10.1007/s11251-013-9285-2
- Kang, S., Lindsey, R., Mozer, M., & Pashler, H. (2014). Retrieval practice over the long term: Should spacing be expanding or equal-interval? *Psychonomic bulletin & review, 21*(6), 1544-1550.
- Karpicke, J., & Roediger, H. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 704–719.
- Keppel, G. (1964). Facilitation in short- and long-term retention of paired associates following distributed practice in learning. *Journal of Verbal Learning and Verbal Behavior, 3*, 91-111.
- Keppel, G. (1967). A reconsideration of the extinction-recovery theory. *Journal of verbal learning and verbal behavior, 6*, 476-486.
- King, J. (2002). Using DVD feature films in the EFL classroom. *Computer Assisted Language Learning, 15*(5), 509-523.
- Koirala, C. (2015). The word frequency effect on second language vocabulary learning. In F. Helm, L. Bradley, M. Guarda, & S. Thouëсны (Eds),. *Critical CALL – Proceedings of the 2015 EUROCALL Conference, Padova, Italy*, 318-323.
- Kornell, N. (2009). Optimising Learning Using Flashcards: Spacing Is More Effective Than Cramming. *Applied Cognitive Psychology, 23*, 1297-1317.

- Kosslyn, S., & Smith, E. (2006). *Cognitive psychology: Mind and brain*. Saddle River, NJ: Prentice-Hall Inc.
- Lado, R., Baldwin, B., & Lobo, F. (1967). *Massive vocabulary expansion in a foreign language beyond the basic course: The effects of stimuli, timing and order of presentation*. Washington, DC: US Department of Health, Education and Welfare.
- Laufer, B. (1986). Possible changes in attitude towards vocabulary acquisition research. *IRAL: International Review of Applied Linguistics in Language Teaching*, 24(1), 69.
- Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? (C. Lauren, & M. Nordman, Eds.) *Special language: From humans to thinking machines*, pp. 316–323.
- Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different? *Applied linguistics*, 19(2), 255-271.
- Laufer, B., & Girsai, N. (2008). Form focused instruction in second language vocabulary learning: A case for contrastive analysis and translation. *Applied linguistics*, 29(4), 694-716.
- Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, 54(3), 399–436.  
doi:10.1111/j.0023-8333.2004.00260.x
- Laufer, B., & Nation, I. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16, 307–322.
- Laufer, B., & Paribakht, T. (2000). The relationship between passive and active vocabularies: Effects of language learning context. *Language Learning*, 48(3), 365–391.

- Laufer, B., & Shmueli, K. (1997). Memorizing new words: Does teaching have anything to do with it? *RELC Journal*, 28(1), 89–108.  
doi:10.1177/003368829702800106
- Laufer, B., Elder, C., Hill, K., & Congton, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing*, 21(2), 202–226.  
doi:10.1191/0265532204lt277oa
- Lengyel, Z., & Navracsecs, J. (2007). Second language lexical processes: applied linguistic and psycholinguistic perspectives. *Multilingual Matters*, 23.
- Leroy, G., & Kauchak, D. (2014). The effect of word familiarity on actual and perceived text difficulty. *Journal of the American Medical Informatics Association : JAMIA*, 21(1).
- Lewis (Ed.), M. (2000). *Teaching Collocation: Further Developments in the Lexical Approach* (Vol. 4). Hove, England: Language Teaching Publications.
- Lotfolahi, A., & Salehi, H. (2017). Pacing effects in vocabulary learning: Young EFL learners in focus. *Cogent Education*, 4(1), 1287391.
- Loudon, I. (2008). The use of historical controls and concurrent controls to assess the effects of sulphonamides, 1936–1945. *Journal of the Royal Society of Medicine*, 101(3), 148-155.
- Lowie, W., & Seton, B. (2012). *Essential statistics for applied linguistics*. Macmillan International Higher Education.
- Martella, R., Nelson, R., & Marchand-Martella, N. (1999). *Research Methods: Learning to Become a Critical Consumer*. New York: Allyn & Bacon.
- Mayer, R., & Anderson, R. (1992). The instructive animation: Helping students build connections between words and pictures in multimedia learning. 84(4), 444.

- McGregor, E., Swabey, K., & Pullen, D. (2015). How often do you move? Improving student learning in the elementary classroom through purposeful movement. *Open Journal of Social Sciences*, 3(6), 6.
- McKenzie, R. (2008). Social factors and non-native attitudes towards varieties of spoken English: a Japanese case study. *International journal of Applied linguistics*, 18(1), 63-88.
- McLean, S., Hogg, N., & Rush, T. (2013). Vocabulary learning through an online computerized flashcard site. *JALT CALL Journal*, 9(1), 79-98.
- Meara, P. (1980). Vocabulary Acquisition: A Neglected Aspect of Language Learning. *Language Teaching*, 13(3-4), 221-246.
- Meara, P. (1989). Matrix models of vocabulary acquisition. *AILA Review*, 6, 66-74.
- Meara, P. (1992). *EFL Vocabulary Tests (2nd ed.)*. University College Swansea: Centre for Applied Language Studies.
- Meara, P., Lightbown, P., & Halter, R. (1997). Classrooms as lexical environments. *Language Teaching Research*, 1(1), 28-48.
- Medina, A. (2019). Issues of Depth of Processing and Think-Aloud Reactivity. *The Routledge Handbook of Second Language Research in Classroom Learning: Processing and Processes*.
- Milliner, B. (2013). Using online flashcard software to raise business students' TOEIC scores. *Annual Report of JACET SIG on ESP*, 15, 51-60.
- Milton, J. (2006). Language lite: learning French vocabulary in school. *Journal of French Language Studies*, 16(2), 187-205.
- Milton, J. (2008). Vocabulary uptake from informal learning tasks. *Language Learning Journal*, 36(2), 227-238.
- Milton, J. (2009). *Measuring Second Language Vocabulary Acquisition*. Clevedon, UK: Multilingual Matters.

- Milton, J., & Meara, P. (1995). How periods abroad affect vocabulary growth in a foreign language. *ITL, 107-108*, 17-34.
- Milton, J., & Meara, P. (1998). Are the British really bad at learning foreign languages? *Language Learning Journal, 18*, 68-76.
- Moss, V. (1995). The efficacy of massed versus distributed practice as a function of desired learning outcomes and grade level of the student.
- Moyer, K., & Von Haller Gilmer, B. (1954). The concept of attention spans in children. *The Elementary School Journal, 54*(8), 464-466.
- Nakata, T. (2006). Implementing optimal spaced learning for English vocabulary learning: Towards improvement of the low-first method derived from the reactivation theory. *The JALT Call Journal, 2*(2), 3-18.
- Nakata, T. (2011). Computer-assisted second language vocabulary learning in a paired-associate paradigm: a critical investigation of flashcard software. *Computer Assisted Language Learning, 24*(1), 17-38.
- Nakata, T. (2015). Effects of expanding and equal spacing on second language vocabulary learning: Does gradually increasing spacing increase vocabulary learning? *Studies in Second Language Acquisition, 37*(4), 677-711.
- Nakata, T. (2017). Does repeated practice make perfect? The effects of within-session repeated retrieval on second language vocabulary learning. *Studies in Second Language Acquisition, 39*(4), 653-679.
- Nation, I. (1980). Strategies for receptive vocabulary learning. *RELC Guidelines, 3*, 18-23.
- Nation, I. (1990). *Teaching and learning vocabulary*. New York: Heinle and Hienle.
- Nation, I. (1997). Bringing today's vocabulary research into tomorrow's classrooms. In G.M. Jacobs (ed.). *Language Classrooms of Tomorrow: Issues and Responses RELC Anthology Series, 38*, 170-182.

- Nation, I. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.
- Nation, I. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, 63(1), 59–82.
- Nation, I. (2007). The four strands. *Innovation in Language Learning and Teaching*, 1(1), 1-12.
- Nation, I. (2011). My ideal vocabulary course. In M.H. Chau & I. A. Khairi, (Eds). *Vocabulary learning in the language classroom*.
- Nation, I. (2014). What do you need to know to learn a foreign language. Retrieved from: [http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/foreign-language\\_1125.pdf](http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/foreign-language_1125.pdf).
- Nation, I., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, 31(7), 9-13.
- Nation, I., & Gu, P. (2007). *Focus on Vocabulary*. Sydney: National Centre for English Language Teaching and Research. Macquarie University.
- Nation, I., & Waring, R. (1997). Vocabulary size, text coverage and word lists. *Vocabulary: Description, acquisition and pedagogy*, 14, 6-19.
- Nesselhauf, N., & Tschichold, C. (2002). Collocations in CALL: An investigation of vocabulary-building software for EFL. *Computer Assisted Language Learning*, 15(3), 251–279. doi:10.1076/call.15.3.251.8190
- Parry, K. (1991). Building a Vocabulary Through Academic Reading. *Tesol Quarterly*, 25(4), 629-653.
- Pashler, H., Rohrer, D., Cepeda, N., & Carpenter, S. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review*, 14, 187-193. Retrieved from <http://escholarship.org/uc/item/7mt714qq>



- Phillips, D., & Siegel, H. (2013). Philosophy of Education. *The Stanford Encyclopedia of Philosophy*, 1.
- Pigada, M. &. (2006). Vocabulary acquisition from extensive reading: a case study. *Reading in a Foreign Language*, 18(1), 1-28.
- Pimsleur, P. (1967). A memory schedule. *The Modern Language Journal*, 51(2), 73-75.
- Pintrich, P., & Schunk, D. (2002). *Motivation in Education: Theory, Research and Applications* (2 ed.). Upper Saddle River, NJ: Merrill Prentice-Hall.
- Porte, G. (Ed.). (2012). *Replication research in applied linguistics*. Cambridge University Press.
- Purnama, A. (2017). Incorporating memes and instagram to enhance student's participation. *LLT Journal: A Journal on Language and Language Teaching*, 20(1), 1-14.
- Pyc, M., & Rawson, K. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, 60(4), 437-447.  
doi:10.1016/j.jml.2009.01.004
- Rasinski, T., Padak, N., & Newton, J. (2017). The roots of comprehension. *Literacy*, 74(5).
- Real Academia Española. (n.d.). *CREA*. Retrieved 2016, from Diccionario de la lengua española: [http://corpus.rae.es/frec/CREA\\_total.zip](http://corpus.rae.es/frec/CREA_total.zip)
- Rohrer, D., & Pashler, H. (2007). Increasing retention without increasing study time. *Current Directions in Psychological Science*, 16(4), 183-186.
- Schmidt, R., & Bjork, R. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological science*, 3(4), 207-218.

- Schmitt, N. (2008). Instructed second language vocabulary learning. *Language teaching research*, 12(3), 329-363.
- Schmitt, N. (2019). Understanding vocabulary acquisition, instruction, and assessment: A research agenda. *Language Teaching*(52), 2.
- Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, 95(1), 26-43.
- Seabrook, R., Brown, G., & Solity, J. (2005). Distributed and Massed Practice: From Laboratory to Classroom. *Applied Cognitive Psychology*, 19, 107–122.
- Serrano, R., & Huang, H. (2018). Learning Vocabulary Through Assisted Repeated Reading: How Much Time Should There Be Between Repetitions of the Same Text? *TESOL Quarterly*, 52(4), 971-994.
- Sherman, J. (2003). *Using authentic video in the language classroom*. Ernst Klett Sprachen.
- Shuell, T., & Lee, C. (1976). *Learning and instruction*. Monterey, California: Brooks/Cole Pub. Co.
- Siegel, M., & Misselt, A. (1984). Adaptive feedback and review paradigm for computerbased drills. *Journal of Educational Psychology*, 76(2), 310–317.
- Singleton, D. (1999). *Exploring the second language mental lexicon*. Ernst Klett Sprachen.
- Slamecka, N., & Graf, P. (1978). The generation effect: delineation of a phenomenon. *Journal of experimental Psychology: Human learning and Memory*, 592.
- Smith, E., & Kosslyn, S. (2007). *Cognitive psychology : Mind and brain / Edward E. Smith and Stephen M. Kosslyn; with the contributions of Lawrence W. Barsalou ... [et al.]*. Upper Saddle River, NJ: Pearson/Prentice Hall.
- Sobel, H., Cepeda, N., & Kapler, I. (2011). Spacing effects in real-world classroom vocabulary learning. *Applied Cognitive Psychology*, 25(5), 763-767.

- Spiri, J. (2008). Online study of frequency list vocabulary with the WordChamp website. *Reflections on English Language Teaching*, 7(1), 21-36.
- Stutz, H. (1992). Flashcards: fast and fun. *Hispania*, 75(5), 1323-1325.
- Suldo, S., Shaunessy, E., & Hardesty, R. (2008). Relationships among stress, coping, and mental health in high-achieving high school students. *Psychology in the Schools*, 45, 273-290.
- Suldo, S., Shaunessy, E., Thalji, A., Michalowski, J., & Shaffer, E. (2009). Sources of stress for students in high school college preparatory and general education programs: group differences and associations with adjustment. *Adolescence*, 44(176), 925-948.
- Sunderman, G., & Kroll, J. (2006). First language activation during second language lexical processing. *Studies in Second Language Acquisition*, 28, 387-422.
- Suzuki, Y., & DeKeyser, R. (2017). Effects of distributed practice on the proceduralization of morphology. *Language Teaching Research*, 21(2), 166-188.
- Swain, M. (1995). Three functions of output in second language learning. In G. Cook and B. Seidelhofer (eds). *Principle and Practice in Applied Linguistics: Studies in Honour of H.G. Widdowson*, 125-144.
- Tüysüz, M., Yildiran, D., & Demirci, N. (2010). What is the motivation difference between university students and high-school students? *Procedia Social and Behavioral Sciences*, 2, 1543-1548.
- Tang, E., & Nesi, H. (2003). Teaching vocabulary in two Chinese classrooms: Schoolchildren's exposure to English words in Hong Kong and Guangzhou. *Language Teaching Research*, 7(1), 65-97.
- Thorndike, E. (1908). Memory for paired associates. *Psychological Review*, 15(2), 122-138. doi:10.1037/h0073570

- Underwood, B. (1961). Ten years of massed practice on distributed practice. *Psychological Review*, 68, 229-247.
- Underwood, B., & Ekstrand, B. (1967). Studies of distributed practice: XXIV. Differentiation and proactive inhibition. *Journal of Experimental Psychology*, 74, 574-580.
- van Bussel, F. (1994). Design rules for computer-aided learning of vocabulary items in a second language. *Computers in Human Behavior*, 10(1), 63–76.  
doi:10.1016/0747-5632(94)90029-9
- Verkoeijen, P., Rikers, R., & Özsoy, B. (2008). Distributed rereading can hurt the spacing effect in text memory. *Applied Cognitive Psychology*, 22, 685–695.
- Vermeer, A. (2004). The relation between lexical richness and vocabulary size in Dutch L1 and L2 children. In P. Bogaards and B. Laufer (eds). *Vocabulary in a Second Language*, 173-189.
- Webb, S. (2009). The effects of pre-learning vocabulary on reading comprehension and writing. *Canadian Modern Language Review*, 65(3), 441–470.  
doi:10.1353/cml.0.0046
- Webb, S., & Piasecki, A. (2018). Re-examining the effects of word writing on vocabulary learning. *ITL-International Journal of Applied Linguistics*, 169(1), 72-94.
- Wells, J. (2011). International education, values and attitudes: A critical analysis of the International Baccalaureate (IB) Learner Profile. *Journal of Research in International Education*, 10(2), 174-188.
- West, M. (1953). *A general service list of English words*. London: Longman, Green.
- Wilkins, D. (1972). *Linguistics in language teaching*. E. Arnold, 1973.

Wilson, W. (1976). Developmental changes in the lag effect: an encoding hypothesis for repeated word recall. *Journal of Experimental Child Psychology*, 22, 113 – 122.

WordEngine. (n.d.). Retrieved from WordEngine: <https://www.wordengine.jp/>

Yashima, T. (2009). International posture and the ideal L2 self in the Japanese EFL context. In T. Taguchi, M. Magid, & M. Papi, *The L2 motivational self system among Japanese, Chinese and Iranian learners of English: A comparative study. Motivation, language identity and the L2 self* (Vol. 36, pp. 66-97).

Zanón, N. (2006). Using subtitles to enhance foreign language learning. *Porta Linguarum: revista internacional de didáctica de las lenguas extranjeras*(6), 4.

## Appendix I

Below there is a screenshot of the confidence test of the replication study.

¿ Cómo te llamas?

Grade

---

1- I am good at Spanish

Strongly Agree  Somewhat Agree  Somewhat Disagree  Strongly Disagree

2- I am good at understanding authentic Spanish materials

Strongly Agree  Somewhat Agree  Somewhat Disagree  Strongly Disagree

3- I am good at understanding videos in authentic Spanish

Strongly Agree  Somewhat Agree  Somewhat Disagree  Strongly Disagree

4- I can learn new words easily

Strongly Agree  Somewhat Agree  Somewhat Disagree  Strongly Disagree

5- I am good at learning new vocabulary from videos in authentic Spanish

Strongly Agree  Somewhat Agree  Somewhat Disagree  Strongly Disagree

---

Submit

## Appendix II

Below there is a screenshot of the pre-test of the replication study.

¿ Cómo te llamas?

Grado

---

1- Un ----- es una ciudad pequeña.  
 fantasma  mundo  pueblo

2- El panda es un animal muy -----  
 tranquilo  gigante  maníaco


3- Los niños tienen ----- a los monstruos  
 miedo  pequeños  aman

4- Los monstruos están ----- en los armarios  
 malos  escondidos  gigantes

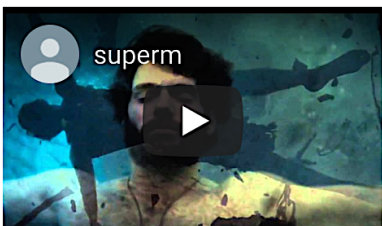
5- La ----- es lo opuesto a la oscuridad  
 magia  luz  poder

6- Un héroe es también -----  
 valiente  princesa  antiguo

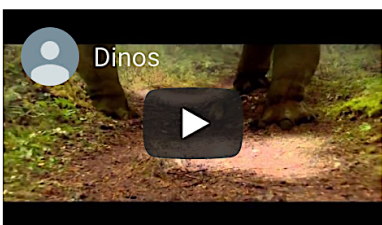
7- El ----- es un insecto verde muy pequeño  
 grulla  víbora  mantis

 Iron3

15- ¿Lo opuesto de vida es...?

 superm

16- ¿Cómo se dice 'world' en español?

 Dinos

17- ¿Cuál es el sinónimo de grande?

8- ¿Cómo se dice 'nice' en español?  
 enorme  antiguo  simpático

9- ¿Cómo se dice 'joke' en español?  
 miedo  fantasma  broma

10- ¿Cómo se dice 'save' en español?  
 atacar  salvar  detener

11- ¿Cómo se dice 'attack' en español?  
 venganza  atacar  morir

12- ¿Cómo se dice 'something' en español?  
 algo  don  mono

13- ¿Cómo se dice 'crane' en español?  
 víbora  mantis  grulla

Mirar los videos y responder las preguntas:



14- ¿Qué animal se nombra en el video?



18- ¿Cómo se dice 'gift' en español?



19- ¿Cómo se dice 'princess' en español?



20- ¿Cuál es el sinónimo de 'me gusta'?



## Appendix III

### TRAILER

### RETRIEVED FROM

---

Man of steel	<a href="http://www.youtube.com/embed/IURsJgR3tfl?rel=0">www.youtube.com/embed/IURsJgR3tfl?rel=0</a>
Iron man 3	<a href="http://www.youtube.com/embed/Ga817IEqAoI?rel=0">www.youtube.com/embed/Ga817IEqAoI?rel=0</a>
Monsters University	<a href="http://www.youtube.com/embed/mSFoeqwXnw4?rel=0">www.youtube.com/embed/mSFoeqwXnw4?rel=0</a>
Walking with dinosaurs	<a href="http://www.youtube.com/embed/nCL3--VQCN4?rel=0">www.youtube.com/embed/nCL3--VQCN4?rel=0</a>
Jack the giant slayer	<a href="http://www.youtube.com/embed/wRZvaPeo63A?rel=0">www.youtube.com/embed/wRZvaPeo63A?rel=0</a>
Paranorman	<a href="http://www.youtube.com/embed/zZI2Z1cqbY4?rel=0">http://www.youtube.com/embed/zZI2Z1cqbY4?rel=0</a>
Kung-fu Panda	<a href="http://www.youtube.com/embed/V8-t-40eMQE?rel=0">http://www.youtube.com/embed/V8-t-40eMQE?rel=0</a>

---

## Appendix IV

The picture below shows a screenshot of one of the readings in the replication study. The words in blue are the target words, which are underlined showing that if the mouse cursor hovers on top of them, there appears the translation into English for that word. Below the reading there are ten multiple-choice questions. After the participant submits the answers, a blue section on the right shows which answers were correct. Once all ten questions were answered to correctly, the movie trailer for that reading is shown at the bottom of the page.

## Caminando con dinosaurios

Patchi es **simpático** y le gustan las **bromas**. El es el **más pequeño** de la **manada** de paquirinosaurios. Los paquirinosaurios eran animales **antiguos enormes** en un **mundo** con animales **gigantes**. Patchi tiene un hermano más **grande** llamado Scowler. Su padre se llama Bulldust y es el líder de la manada. En un **viaje épico**, un dinosaurio llamado Gorgon **ataca** a la manada, y Bulldust **muere**. Los hermanos son **supervivientes**, y abandonan la manada. Años después, Patchi y Scowler **regresan** al grupo. Scowler es el líder y Patchi abandona la manada. Al final de la película Patchi es muy **valiente**, regresa para **proteger** y **salvar** a su familia y se **convertirá** en héroe.

### Responder las preguntas

- 1- El capitán hizo un ---- muy largo en el océano.  
 viaje  broma  robot
- 2- Los dinosaurios son ----.  
 pequeños  gigantes  escondidos
- 3- Los ---- no están muertos.  
 manada  muertos  supervivientes
- 4- Grandioso es un sinónimo de ----.  
 épico  malo  viajes
- 5- El sinónimo de grande es ----.  
 enorme  pequeño  malo
- 6- Los niños se ---- en hombres.  
 duermes  convierten  aprender
- 7- Superman es un ----.  
 gigante  héroe  maníaco
- 8- El dinosaurio es ---- grande que un perro.  
 más  pequeño  gigante
- 9- Un ---- no tiene miedo.  
 enorme  elegante  valiente
- 10- Hay muchos animales en una ----.  
 enorme  viaje  manada

Submit

### RESPUESTAS:

- Respuesta: 1: *viaje* es Correcta.
- Respuesta: 2: *gigantes* es Correcta.
- Respuesta: 3: *supervivientes* es Correcta.
- Respuesta: 4: *épico* es Correcta.
- Respuesta: 5: *enorme* es Correcta.
- Respuesta: 6: *convierten* es Correcta.
- Respuesta: 7: *héroe* es Correcta.
- Respuesta: 8: *más* es Correcta.
- Respuesta: 9: *valiente* es Correcta.
- Respuesta: 10: *manada* es Correcta.
- Total respuestas correctas: 10



## Appendix V

The instructions below were read at the beginning of the intervention by the teachers in charge of each group in the replication study.

----- Here below the actual text given to students -----

You are about to start working on a research project that investigates how students can improve their comprehension of movie trailers and their retention of a number of Spanish words. The project will take place in four lessons over a period of two weeks. You are required to work individually on your personal computer focusing strictly on the activities assigned. You will be given a link to a test that will ask questions regarding how confident you feel when working with authentic materials. The next activity will be another test where you need to answer questions and in some other cases you have to watch a video clip and answer questions about it. Later you will work on readings that have some words in blue. You can bring the mouse cursor of your computer over the blue words and an English translation will be provided. Below the reading you will find a series of questions. After you have answered all questions correctly a movie trailer will appear at the bottom of the page and you will be able to watch it. Continue to follow the links provided and follow the instructions of your teacher along the duration of the project. You are expected to work on your own, working on the activities to the best of your abilities. Your performance in this project will not affect your Spanish course grade.

## Appendix VI

The table below shows all of the originally pre-selected 120 words and their score from the target vocabulary pre-selection test. The aim of the test was to elicit the 100 words participants did not know to be used as target words in the project. Therefore, every time a participant answered correctly to the meaning of a word, that word received one point. The words with the highest points were the words that participants answered correctly to the most and were excluded from the final 100 target word list. Scores below show that the words *aburrido*, *mirar*, and *moreno* had a similar score of eight (8). Therefore, after checking test scores, the researcher decided to test each one of the participants orally on the meaning of those three words. The word that most of the participants answered wrongly to in the second test was: *aburrido*, which was finally included in the final list of target words.

abrir, to open	11	cómodo, comfortable	5	jugo, juice	6	precio, price	7
aburrido, boring*	8	compañero, partner	4	lago, lake	5	periodista, journalist	4
adolescente, teenager	4	comprar, to buy	6	libre, free	6	primo, cousin	3
agua, water	11	conducir, to drive	4	llover, to rain	5	regresar, to return	5
alegre, happy	6	correo electrónico, email	10	luna, moon	7	revista, magazine	5
almuerzo, lunch	6	cuaderno, notebook	7	mes, month	7	rico, rich	13
alquilar, to rent	3	cubiertos, silver-ware	4	mientras, while	4	ropa, clothes	11
alumno, student	5	cuerpo, body	5	mirar, to look at*	8	roto, broken	0
árbol, tree	10	cuidar, look after	6	moda, fashion	6	saber, to know something	6
asado, barbeque	3	curar, to cure	6	moneda, coin	6	salvar, to save	5
asignatura, subject	11	decir, to say	5	moreno, dark*	8	secundaria, high-school	7
ayer, yesterday	6	demasiado, too many	4	necesitar, to need	10	semana, week	5
ayudar, to help	5	después, after	5	nieve, snow	2	simpático, nice	9
beber, to drink	11	dibujo, drawing	3	nube, cloud	7	significar, to mean	4
biblioteca, library	7	dinero, money	10	nublado, cloudy	4	sol, sun	6
billete, ticket	4	divertido, fun	4	odiar, to hate	6	soleado, sunny	3
bolsa, bag	1	dolor, pain	2	pájaro, bird	5	solicitar, to request	5
cabello, hair	5	dueño, owner	4	pan, bread	11	sueldo, salary	4
calle, street	4	durante, during	6	panadería, bakery	5	tamaño, size	4
camarero, waiter	2	edad, age	10	pariente, relative	3	también, also	9
cambiar, to change	4	enfermero, nurse	4	película, movie	5	tampoco, either	3
caminar, to walk	7	equipo, team	3	peligroso, dangerous	6	tarjeta, card	6
carne, meat	1	extranjero, foreigner	4	perder, to lose	5	techo, roof	2
carta, letter	9	folleto, pamphlet	6	perezoso, lazy	9	trabajo, job	6
casado, married	2	guerra, war	3	periódico, newspaper	5	vaso, glass	3
cena, dinner	3	hacer, to do	4	periodista, journalist	7	viajar, to travel	7
centro comercial, mall	7	herida, wound	2	pero, but	11	viaje, journey	7
cielo, sky	6	horario, schedule	2	pescado, fish	5	video juego, video game	6
cita, appointment	4	hoy, today	7	pierna, leg	7	viejo, old	7
coche, car	11	joven, young	5	piscina, swimming-pool	6		
comida, food	6			plato, plate	6		

## Appendix VII

This appendix provides a small sample of how vocabulary was presented explicitly to students during the topic *el individuo* ('the individual').

----- Here below the actual text given to students -----

### Descripción física

¿Cómo eres?

¿Cómo es tu hermano?

yo soy	yo tengo	yo llevo
tu eres	tu tienes	tu llevas
él/ella es	él/ella tiene	él/ella lleva
pelo corto	ojos verdes	alto/a
pelo largo	ojos marrones	bajo/a
pelo liso	ojos azules	guapo/a
pelo ondulado	ojos negros	feo/a
pelo rizado	gafas	delgado/a
pelo rubio	barba	gordo/a
pelo castaño	bigote	joven
pelo marrón	pecas	viejo/a
pelo gris		
pelo negro		
calvo/a		

## Appendix VIII

The table below shows the final list of 100 words used as target words in the study.



aburrido, boring	comprar, to buy	durante, during	nieve, snow	roto, broken
adolescente, teenager	conducir, to drive	enfermero, nurse	nube, cloud	saber, to know something
alegre, happy	cuaderno, notebook	equipo, team	nublado, cloudy	salvar, to save
almuerzo, lunch	cubiertos, silver-ware	extranjero, foreigner	odiar, to hate	secundaria, high-school
alquilar, to rent	cuerpo, body	folleto, pamphlet	pájaro, bird	semana, week
alumno, student	cuidar, look after	guerra, war	panadería, bakery	significar, to mean
asado, barbeque	curar, to cure	hacer, to do	pariente, relative	sol, sun
ayer, yesterday	decir, to say	herida, wound	película, movie	soleado, sunny
ayudar, to help	demasiado, too many	horario, schedule	peligroso, dangerous	solicitar, to request
biblioteca, library	después, after	hoy, today	perder, to lose	suelo, salary
billete, ticket	dibujo, drawing	joven, young	periódico, newspaper	tamaño, size
bolsa, bag	divertido, fun	jugo, juice	pescado, fish	tampoco, either
cabello, hair	centro comercial, mall	lago, lake	pierna, leg	tarjeta, card
calle, street	cielo, sky	libre, free	piscina, swimming-pool	techo, roof
camarero, waiter	cita, appointment	llover, to rain	plato, plate	trabajo, job
cambiar, to change	comida, food	luna, moon	precio, price	vaso, glass
caminar, to walk	cómodo, comfortable	mes, month	periodista, journalist	viajar, to travel
carne, meat	compañero, partner	mientras, while	primo, cousin	viaje, journey
casado, married	dolor, pain	moda, fashion	regresar, to return	video juego, video game
cena, dinner	dueño, owner	moneda, coin	revista, magazine	viejo, old

## Appendix IX

The three lists below were given to participants to learn about Spanish simple past tense.

----- Here below the actual text given to students -----

### Presente

#### Verbos especiales

	<b>ar</b>	<b>er</b>	<b>ir</b>	<b>ser</b>	<b>hacer</b>	<b>ir</b>	<b>gustar</b>
Yo	o	o	o	soy	hago	voy	me gusta
Tu	as	es	es	eres	haces	vas	te gusta
El	a	e	e	es	hace	va	le gusta
Nos.	amos	emos	imos	somos	hacemos	vamos	nos gusta
Vos.	áis	éis	ís	sois	hacéis	vais	os gusta
Ellos	an	en	en	son	hacen	van	les gusta

## Pasado

### Verbos especiales

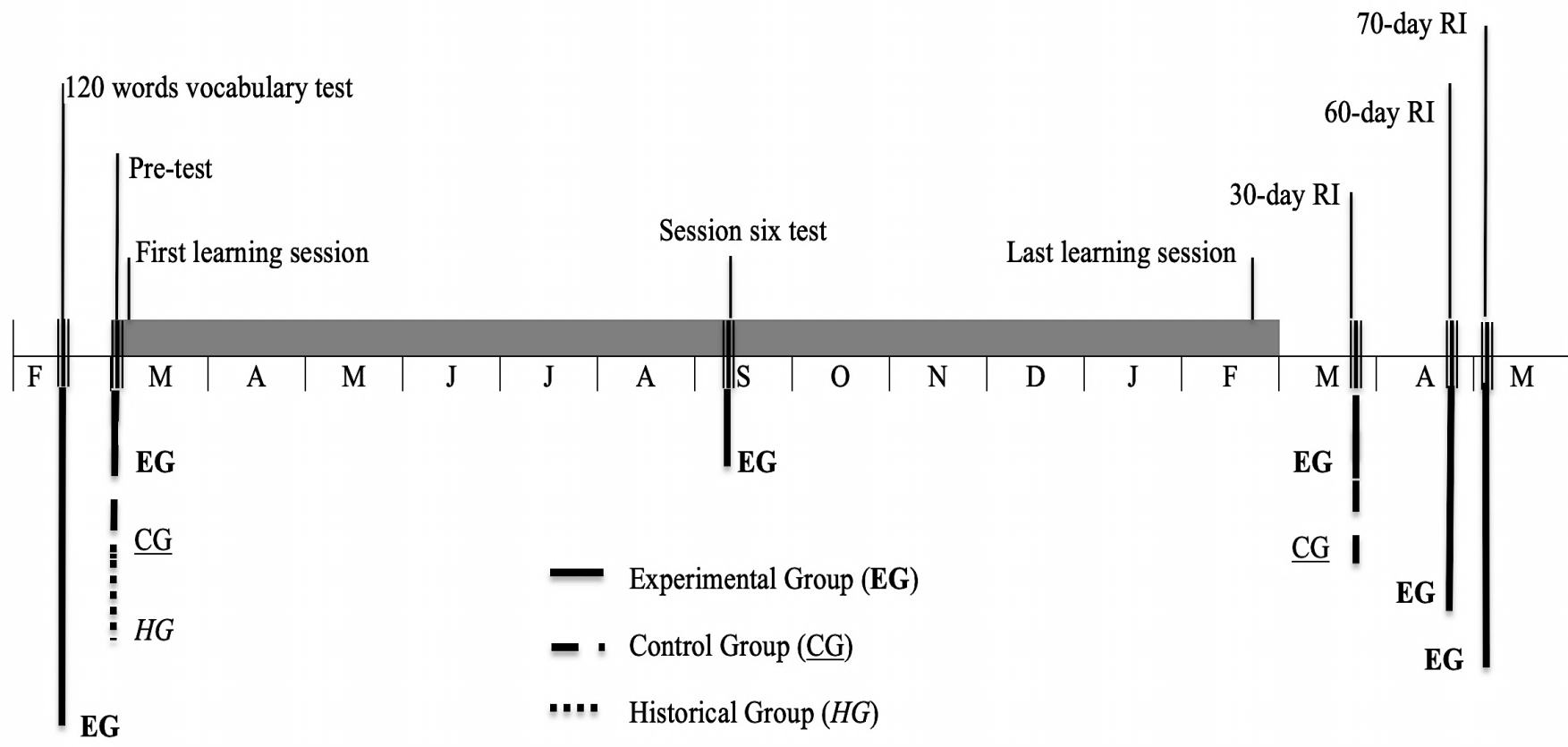
	<b>ar</b>	<b>er</b>	<b>ir</b>	<b>ser</b>	<b>hacer</b>	<b>ir</b>	<b>gustar</b>
Yo	é	í	í	fui	hice	fui	me gustó
Tu	aste	iste	iste	fuiste	hiciste	fuiste	te gustó
El	ó	ió	ió	fue	hizo	fue	le gustó
Nos.	amos	imos	imos	fuimos	hicimos	fuimos	nos gustó
Vos.	asteis	isteis	ísteis	fuisteis	hicistéis	fuisteis	os gustó
Ellos	aron	ieron	ieron	feron	hicieron	fueron	les gustó

### Algunos verbos extra

vivir	tocar	cortar	despertarse
estudiar	leer	lavar	levantarse
trabajar	comprar	pelar	ducharse
tener	gastar	freír	levarse
comer	ahorrar	poner	bañarse
desayunar	ver	mezclar	vestirse
cenar	limpiar	batir	acostarse
cocinar	aprender	añadir	
escuchar	cantar	sacar	
nadar	bailar	calentar	
jugar	viajar	servir	
hacer	ir	alojarse	
llevar	caminar		
pasear	pasar		
salir	conocer		
donar	tomar		
tocar	disfrutar		
ayudar	dormir		

## Appendix X

The detailed timeline below shows project's development by month (starting in February and finishing in May the following calendar year), together with learning sessions, all tests, and when each group took them. Notice that the experimental group (EG) took six tests overall, the control group (CG) took two, and the historical group (HG) took only one. The post-tests are shown according to their number of days (RI) counting from the last learning session. Notice that the upper section of the timeline shows the tests and learning sessions, while the lower part shows which group was involved in that test or session.



## Appendix XI

This appendix presents the ethics procedure that was employed in the study. After submission of the appropriate documentation, this research project was approved by the Swansea University Research Committee.

----- Here below the actual text given to students -----

Dear Guardians,

Your son/daughter has been asked to participate in a research project led by their Spanish Ab Initio teacher as part of his PhD course. The project will investigate retention of Spanish vocabulary over extended periods of time. The project will take place at the school, during regular school hours so no extraordinary accommodations are necessary on your side. School administration has been duly notified and they support the study. Participation in the project is optional, so be ensured that your child will not be forced to participate, grades will not be affected, and school rules and local customs and traditions will be respected as usual. At the same time, the researcher will also ensure that students' attention is not deviated from their regular course of studies as a result of this research project.

If you consider that your child should not participate please inform the researcher by sending a written note expressing so. Should you have any questions or concerns, please do not hesitate to ask.

Sincerely

## Appendix XII

### Pre Test

**In this test you are asked to answer 30 multiple choice questions.**

**You cannot use any online or paper resources to help you.**

**Please take your time to think about your answers as you can try only once.**

**You have 30 minutes to finish.**

1. La enfermera cuida al /// que no tiene parientes. (1 point)

- dolor
- viejo
- viaje

2. No me gusta la revista, ni /// el periódico. (1 point)

- demasiado
- durante
- tampoco

3. El periodista /// mas sueldo en la revista. (1 point)

- solicita
- precio
- significa

4. Necesito un vaso y /// para comer la cena. (1 point)

- camarero
- cubiertos
- almuerzo



5. El dueño /// una casa muy cómoda con techo rojo. (1 point)

- compro
- alquila
- cambiamos

6. Mientras /// miro una película en mi casa. (1 point)

- llueve
- divertida
- nublado

7. La /// adolescente es muy divertida y colorida. (1 point)

- moda
- video juego
- folleto

8. Hoy /// mi billete de bus para una semana después. (1 point)

- mes
- libre
- cambio

9. Tengo un folleto de un /// con nieve. (1 point)

- piscina
- jugo
- lago

10. Hay muchas nubes, no está ///. (1 point)

- nublado
- sol
- soleado

11. El /// del periódico está mas alto. (1 point)

- billete
- precio
- moneda

12. El camarero /// con un plato con pescado y un jugo. (1 point)

- sabe
- regresa
- comida

13. Siempre /// mis dibujos para arte en la biblioteca. (1 point)

- alumno
- ayuda
- hago

14. Mi primo /// un automovil importado. (1 point)

- conduce
- peligroso
- dueño

15. Ayer /// por la calle muy aburrido. (1 point)

- perdió
- camino
- caminé

16. Siempre compro la comida en la ///. (1 point)

- cena
- almuerzo
- panadería

17. Mario /// el tamaño de la tarjeta de crédito. (1 point)

- odia
- solicito
- salvo

18. El joven extranjero tiene una /// de guerra. (1 point)

- compañero
- herida
- trabajo

19. El niño hace un /// divertido con muchos colores. (1 point)

- alegre
- cuadernos
- dibujo

20. El alumno /// no estudia demasiado en la escuela secundaria. (1 point)

- horario
- video juegos
- adolescente

21. El /// cura el cuerpo del paciente en el hospital. (1 point)

- periodista
- enfermero
- extranjero

22. Siempre /// a mi amigo con su cuaderno del colegio. (1 point)

- regresa
- digo
- ayudo

23. Tengo una /// rota, necesito una cita con el médico. (1 point)

- pierna
- cuerpo
- curar

24. Ayer /// mucho en el centro comercial. (1 point)

- tarjeta
- compré
- bolsa

25. En la montaña hace frío y hay mucha /// blanca. (1 point)

- nieve
- luna
- nubes

26. Siempre como carne asada para la ///. (1 point)

- almuerzo
- cena
- comida

27. /// de estudiar siempre como la cena. (1 point)

- después
- mientras
- durante

28. El /// de fútbol regresa al estadio. (1 point)

- video juego
- jugo
- equipo

29. De noche, en el cielo hay una /// muy grande. (1 point)

- sol
- luna
- soleado

30. Me gusta la carne, mi favorito es el ///. (1 point)

- asado
- casado
- cena

## Appendix XIII

This multiple-choice test was administered online in session four and ten. The Spanish words selected for this test were the ones considered by participants as still unknown to them during sessions one to three.

<b>Spanish</b>	<b>Answer</b>	<b>Distractor A</b>	<b>Distractor B</b>
cubiertos	silver-ware	dinner	hair
aburrido	boring	happy	to walk
adolescente	teenager	street	nurse
camarero	waiter	bag	war
cómodo	comfortable	team	today
después	after	young	to do
alumno	student	foreigner	during
cuaderno	notebook	mall	yesterday
ayudar	to help	meat	food
compañero	partner	schedule	pain
dueño	owner	ticket	dinner
dibujo	drawing	body	street
cambiar	to change	married	bag
almuerzo	lunch	to rent	team
cita	appointment	to say	happy
asado	barbeque	fun	young
curar	to cure	moon	foreigner
herida	wound	son	mall

conducir	to drive	plate	meat
demasiado	too many	free	schedule
folleto	pamphlet	juice	ticket
cuidar	look after	cousin	body
comprar	to buy	cloud	married
divertida	fun	dangerous	to rent
tamaño	size	coin	to say
película	movie	glass	hair
pájaro	bird	swimming-pool	to walk
panadería	bakery	week	nurse
periódico	newspaper	to return	war
pescado	fish	price	today
tarjeta	card	job	to do
llover	to rain	snow	during
perder	to lose	lake	yesterday
pariente	relative	broken	food
secundaria	secondary	month	pain
mientras	while	journey	dinner
pierna	leg	fashion	street
periodista	journalist	magazine	bag
sueldo	salary	video games	team
significar	to mean	sunny	happy
odiar	to hate	hair	young
solicitar	to request	to walk	foreigner
biblioteca	library	nurse	mall
saber	to know how to	war	meat

viajar	to travel	today	schedule
nublado	cloudy	to do	ticket
salvar	to save	during	body
tampoco	either	yesterday	married
techo	roof	food	to rent
viejo	old	pain	to say

## Appendix XIV

Reading activity for learning session three.

----- Here below the actual text given to students -----

### READING AND MEMES

Manuel es un joven adolescente alumno en una escuela secundaria internacional en Lima, Perú. El chico dice que la escuela no es divertida, él está siempre aburrido en todas las asignaturas, y tiene un horario muy ocupado. El joven prefiere estar en su casa porque está más cómodo y le encanta jugar a los video juegos. El no está muy alegre mientras está en la escuela, y sus compañeros tampoco están contentos con él.

Las asignaturas favoritas de Manuel en el colegio son dibujo y economía. No le gusta francés porque no sabe qué significan las palabras.

Su cuaderno de arte tiene muchos dibujos muy bonitos, y folletos. El hace dibujos soleados muy interesantes con mucho sol, pájaros libres en el cielo azul, lagos grandes con muchos pescados, y muchas plantas. Los folletos tienen calles de una ciudad con un clima muy nublado, con demasiadas nubes, sin luna. Tienen mucha nieve, y también está por llover.

A Manuel le gusta mucho el fútbol. El equipo favorito de Manuel es el Barcelona. Manuel odia a su equipo cuando ellos pierden.

El padre del joven Manuel se llama Ricardo. El es un viejo de cabello blanco que tiene un trabajo como periodista en un periódico. A Ricardo le gusta mucho leer libros y revistas y está siempre en una biblioteca. El señor trabaja mucho en la semana, pero recibe un sueldo muy malo al mes.

Para el desayuno a Ricardo le gusta un vaso de jugo de tomate. Su comida favorita es la carne y prefiere el asado para el almuerzo y come con cubiertos muy elegantes.

Durante la cena Ricardo come un plato grande de vegetales. A la noche le gusta mucho mirar películas del ayer donde salvan a gente en guerras peligrosas.

La madre de Manuel se llama Noelia y ella es dueña de una panadería. A Noelia le gusta conducir su auto pequeño para ir al centro comercial. Ella compra pantalones de moda de mucho precio y prefiere utilizar la tarjeta de crédito. Siempre regresa a la casa muy alegre.

Manuel tiene muchos parientes. La prima de Manuel se llama Cristina. Ella está casada con un señor extranjero y trabaja de camarera. Ella alquila una casa de techo rojo, y piscina de gran tamaño en el centro de la ciudad. A Cristina le gusta mucho viajar. Ella tiene una bolsa con muchas monedas y billetes de países diferentes, y siempre cambia monedas para coleccionar.

Hoy Cristina no va a trabajar porque no está bien. Cristina está herida, tiene una herida en el cuerpo, y una pierna rota y no puede caminar. Hoy ella solicita una cita con el médico para que cure su herida. Después va a necesitar ayuda con el dolor. El enfermero la va a cuidar.



## Appendix XV

The table below shows all of the target words and the percentage of participants (N 15) answering correctly to them in the pre-test and in the test in session six. The first 15 words listed in each column are the ones participants missed the most in both tests and were used for the activity in session seven. Considering most words appeared twice, the complete list for the activity resulted in 17 words (*cena, solícita, viejo, cubiertos, tampoco, herida, regresa, caminé, panadería, equipo, ayudo, odia, llueve, soleado, precio, alquila, hago*).

<b>pre-test</b>		<b>session six</b>	
<b>target word</b>	<b>%</b>	<b>target word</b>	<b>%</b>
cena	13%	regresa	33%
solicita	13%	viejo	33%
viejo	20%	tampoco	33%
cubiertos	20%	solicita	40%
tampoco	20%	herida	40%
herida	27%	cena	40%
regresa	27%	odia	40%
caminé	27%	llueve	47%
panadería	33%	alquila	53%
equipo	33%	precio	53%
ayudo	33%	panadería	53%
odia	33%	hago	53%
llueve	33%	caminé	60%
soleado	40%	soleado	60%
precio	40%	cubiertos	60%
alquila	43%	ayudo	67%
lago	47%	pierna	67%
asado	47%	después	67%
nieve	47%	equipo	73%
moda	53%	lago	73%
compré	53%	asado	73%
hago	60%	moda	73%
pierna	67%	cambio	73%
dibujo	67%	conduce	73%
luna	73%	compré	87%
enfermero	73%	adolescente	87%
cambio	73%	enfermero	87%
conduce	73%	dibujo	93%
adolescente	80%	nieve	100%
después	80%	luna	100%

## Appendix XVI

The screenshot below shows a portion of the fill-in the blank activity in session seven.

**COMPLETAR LA TABLA AZUL CON LA PALABRA CORRECTA.**

**NOTA**

respuesta 1 es CORRECTA

respuesta 2 es incorrecta

respuesta 3 es incorrecta

respuesta 4 es incorrecta

respuesta 5 es incorrecta

respuesta 6 es incorrecta

respuesta 7 es incorrecta

respuesta 8 es incorrecta

respuesta 9 es incorrecta

respuesta 10 es incorrecta

respuesta 11 es incorrecta

respuesta 12 es incorrecta

respuesta 13 es incorrecta

respuesta 14 es incorrecta

respuesta 15 es incorrecta

**-INSTRUCCIONES:**

- Do NOT use Spanish characters (á , é , ñ, etc)

- Click 'tips' below and unscramble the words to be used.

For the list of scrambled words click below:

[Tips](#)

1- boring	aburrido
2- teenager	
3- happy	
4- lunch	
5- to rent	
6- student	
7- barbecue	
8- yesterday	
9- to help	
10- library	
11- ticket	
12- bag	
13- hair	
14- street	

## Appendix XVII

The activity below was assigned to participants in session eight and consisted of fixing mistakes in the sentences provided. All sentences had one mistake and they all used target words in them. Mistakes consisted of wrong word choice based on word meaning. Participants were required to comprehend the meaning of the target words to spot the mistakes and were asked to either change the target word in the sentence, or to make any other necessary changes for the sentence to be grammatically and logically correct.

----- Here below the actual text given to students -----

Instructions: Read the sentences below and correct the mistakes so they are all grammatically and logically correct. All sentences have at least one mistake.

El joven come la cena a la mañana.

El doctor solicita la herida del enfermo.

El viejo es un alumno adolescente.

La señora come los cubiertos en el restaurante.

Me gusta caminar y tampoco odio correr.

La profesora regresa al alumno con la tarea.

El precio de la carne en la panadería es muy alto.

No está soleado, y hay mucho equipo porque llueve mucho.

Mi primo alquila un jugo de naranjas.

Siempre hago a México en las vacaciones.

## Appendix XVIII

This appendix shows all results of the test described in Appendix XII every time it was administered (pre-test, session six, 30-day RI post-test, 60-day RI post-test, and 70-day RI post-test), and for all project groups. However, it is important to remember that (as explained in 4.3.3.3 above) not all groups took all tests.

The test consisted of 30 questions and 30 maximum possible points. The tables below show raw score results for each participant in each test. Notice that Session 6 (in the first table) refers to the test participants in the experimental group took in learning session six. Post-test 30, 60, and 70, refer to the tests taken 30, 60, and 70 days after the last learning session.

While each column of the table represents a different test, rows were discriminated by participants. This means that each row shows results obtained by the same participant in each one of the tests.

All test scores of the experimental group:

<b>Pre-test</b>	<b>Session 6</b>	<b>30-day RI</b>	<b>60-day RI</b>	<b>70-day RI</b>
11	16	24	21	19
10	10	21	22	21
14	18	22	19	18
19	25	27	28	27
13	16	25	22	20
14	25	28	30	28
17	28	28	30	29
12	17	22	20	18
10	13	25	23	22
15	18	28	28	26
13	10	27	23	20
16	21	29	29	27
12	17	19	21	19
15	23	28	27	25
11	16	16	19	17

All test scores of the control group:

<b>Pre-test</b>	<b>30-day RI</b>
10	14
6	11
12	15
8	21
17	17
11	20
12	15
18	21
12	12
13	17
11	14
16	27

Test scores of the historical group

**Pre-test**

---

17

19

14

19

13

15

11

6

19

12

28

26

18

9

14

16

13

12

22

20

19

17

16

16

15

7

12

22

12

18

17

13

---



## Appendix XIX

The table below shows raw score results for each participant in each test in session four and ten.

While each column of the table represents a different test, rows were discriminated by participants. This means that each row shows results obtained by the same participants in each one of the tests.

<b>Session 4</b>	<b>Session 10</b>
45	43
46	50
42	46
48	50
45	50
50	50
50	49
47	49
47	50
46	50
41	48
50	50
49	50
47	49
33	46