TESTING THE REMINDING ACCOUNT OF THE LAG EFFECT IN L2 VOCABULARY ACQUISITION FROM L2-L1 RETRIEVAL PRACTICE WITHIN A PAIRED-ASSOCIATE LEARNING FORMAT

By

Natalya G Koval

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Second Language Studies – Doctor of Philosophy

2020

ABSTRACT

TESTING THE REMINDING ACCOUNT OF THE LAG EFFECT IN L2 VOCABULARY ACQUISITION FROM L2-L1 RETRIEVAL PRACTICE WITHIN A PAIRED-ASSOCIATE LEARNING FORMAT

By

Natalya G Koval

The spacing/lag effect refers to the finding in memory research that spacing repeated study more widely produces important learning benefits (Crowder, 1976; Dempster, 1988, 1989). In order to know when and how this effect can be most useful for second language learning, it is important to understand the cognitive mechanism(s) that drive any effects of spacing in second language learning. It is also important to understand how the operation of the mechanism(s) may be affected by variables inherent in second language learning contexts. In the present study, I investigate the contribution of the dual mechanism of effortful successful retrieval to the effects of lag in second language vocabulary learning. This dual mechanism is proposed to underlie both beneficial and detrimental effects of lag on learning within the reminding account (Benjamin & Tullis, 2010). I additionally investigate the potential effects of externally imposed study time on learning as well as on the operation of the two mechanisms under investigation.

Fifty-two native speakers of American English studied 72 novel L2 Finnish words during overt oral L2-L1 translation retrieval practice in a paired-associate learning format from 6 repetitions under three constant levels of within-session lag with immediate study of feedback for 3 or 9 seconds after each retrieval attempt. Study-phase response latencies and accuracy were recorded and used as measures of study-phase retrieval effort and success, respectively (as in Maddox & Balota, 2015). Immediate and delayed form recognition, L2-L1 translation and translation matching posttests were used to measure learning outcomes.

Results showed a large spacing effect on all measures and at both times of test administration as well as a lag effect on delayed meaning tests. Study time had an overall small positive effect on learning; however, it did not cancel out negative effects of massing retrieval practice: the effects of spacing were considerably larger. Increasing lag between retrieval attempts produced increasingly longer study-phase response latencies and increasingly lower levels of study-phase retrieval success. Study time had a small nonsignificant negative effect on study-phase response latencies and a small significant positive effect on study-phase retrieval success. Moderated mediation analyses showed that study time, as operationalized in the present study, did not affect the operation of the two underlying mechanisms under investigation. They further showed that, despite the fact that a nonmonotonic function was not observed in the present learning outcomes, increasing inter-study interval still had a negative effect on learning and this effect operated through a lower rate of study-phase retrieval success. Further, the moderated mediation analyses showed that the positive effects of retrieval effort (Roediger & Karpicke, 2006) were conditional on retrieval success, in line with predictions of the reminding account.

The findings of the dissertation suggest that: (a) massed L2-L1 translation retrieval practice may not be effective for L2 vocabulary learning; (b) externally imposing a longer study time does not have the large benefits that learner-regulated longer study time does; (c) effortful successful retrieval underlies benefits of lag in L2 vocabulary learning from L2-L1 retrieval practice – the benefits of effortful retrieval are conditional on retrieval success, even in the presence of immediate feedback; (d) successful retrieval is more beneficial than unsuccessful retrieval, even when retrieval attempts are followed by immediate feedback – study of feedback does not offset the negative effects of retrieval failure.

ACKNOWLEDGEMENTS

I owe a tremendous debt of gratitude to my advisors Charlene Polio, Susan Gass, Patti Spinner, Aline Godfroid, and Sandra Deshors for their help, support, and guidance throughout the four years of the SLS program at Michigan State University. These amazing people inspired me and fostered in me a passion for second language research. They provided important guidance and were instrumental in helping me to become the researcher that I am today.

I am further grateful to my professors and classmates for a stimulating experience. I am honored to have met and worked with so many accomplished researchers and simply awesome people! I am grateful to all my participants for taking the time to participate in this study and for making data collection easy and fun.

I am grateful to all my friends outside the SLS program for their friendship and the wonderful time spent together. Finally, I am grateful that I was able to spend these four years living in the breathtakingly beautiful place known as Spartan Village.

Michigan State University has been a great place to study, teach, and do research. My very first experience at MSU had been during the 2000-2001 academic year, when I was an undergraduate exchange student here on an FSA scholarship. I did not know at the time that I would be getting my doctoral degree from this very institution 20 years later! I am honored to call myself a Spartan! Thank you, MSU!

TABLE OF CONTENTS

LIST OF TABLES viii

LIST OF FIGURES x

CHAPTER 1 SECOND LANGUAGE ACQUISITION, OVERT RETRIEVAL PRACTICE, AND THE SPACING/LAG EFFECTS 1

- The present dissertation 3
    - Overview of the dissertation 5
    - Definition of key terms 5

CHAPTER 2 THE SPACING AND LAG EFFECTS AND EFFORTS TO UNDERSTAND THEIR UNDERLYING MECHANISMS 7

- Theories of the spacing effect 9
- The two-process reminding account 14
    - Investigating the role of attention and effort in learning from repetition 15
    - Investigating the role of reminding in learning from repeated study 22
        - Overt retrieval practice and its effects on memory 27
- Research into the spacing effect and retrieval practice in second language acquisition 33
    - Research questions 42

CHAPTER 3 METHOD 43

- Participants 43
- Materials and design 44
    - Study phase 44
    - Distractor math task 48
    - Posttests 49
    - Linguistic background questionnaire 52
- Instruments 53
- Procedure 53
- Analyses 57

CHAPTER 4 RESULTS 58

- Background questionnaire 58
- Posttests results 58
    - Posttest results: Descriptive statistics 59
    - Posttest results: Inferential statistics 69
        - The form-recognition test 70
        - The L2-L1 translation test 74
        - The form-meaning matching test 78
- Study-phase results 82
    - Study-phase response latencies: Descriptive statistics 83
    - Study-phase response latencies: Inferential statistics 89

Study-phase retrieval success: Descriptive statistics 96
Study-phase retrieval success: Inferential statistics 99
Moderated mediation analyses 104
Moderated parallel mediation analyses 104
The form-recognition test 107
The immediate meaning test 109
The delayed meaning test 111
Mediation by retrieval effort moderated by retrieval success (a moderated mediation analysis) 112
The form-recognition scores 115
The immediate meaning scores 116
The delayed meaning scores 117

CHAPTER 5 DISCUSSION 120
Pedagogical implications 128
Limitations and suggestions for future research 132

APPENDICES 137
Appendix A: Target Finnish words with their English translations 138
Appendix B: Information on the English translations 139
Appendix C: Instructions for vocabulary posttests 140
Appendix D: The form recognition test 141
Appendix E: The L2-L1 translation test 145
Appendix F: The form-meaning matching test 147
Appendix G: Linguistic Background Questionnaire.. 149
Appendix H: Study-phase instructions 150

REFERENCES 151

LIST OF TABLES

Table 1: Definitions for key terminology 6

Table 2: Variables used in the study-phase analyses 49

Table 3: Variables used in the posttest analyses 51

Table 4: Variables used in the moderated mediation analyses 52

Table 5: Raw posttest scores in the practice and no-practice conditions 60

Table 6: Raw posttest scores across the three experimental conditions 61

Table 7: Percent correct in the massed practice and no-practice conditions 63

Table 8: Percent correct in the short-spaced practice and no-practice conditions .. 64

Table 9: Percent correct in the long-spaced practice and no-practice conditions ... 65

Table 10: Correlations and loadings for each test on the extracted component 66

Table 11: Form-recognition omnibus test 71

Table 12: Form-recognition results against the no-practice condition 73

Table 13: Form-recognition results against the short-spaced practice condition 74

Table 14: L2-L1 translation omnibus test 76

Table 15: L2-L1 translation results against the no-practice condition 77

Table 16: L2-L1 translation results against the short-spaced practice condition 78

Table 17: Form-meaning matching omnibus test 80

Table 18: Form-meaning matching results against the no-practice condition 81

Table 19: Form-meaning matching results against the short-spaced practice condition 82

Table 20: Response latencies across the practice conditions 84

Table 21: Response latencies in the two study time conditions 85

Table 22: Response latencies in successful and unsuccessful retrieval attempts 86

Table 23: Correct retrieval events per experimental condition 97

Table 24: Study-phase retrieval success in the short and long study time conditions 97

Table 25: The effect of ISI on retrieval success at the five repetitions 102

Table 26: Parameter estimates for the effect of ISI on study-phase retrieval success 103

Table 27: Correlation coefficients and loadings for form-recognition tests 105

Table 28: Correlation coefficients and loadings for immediate meaning tests 105

Table 29: Correlation coefficients and loadings for delayed meaning tests 105

Table 30: Correlation coefficients and loadings for form-recognition tests 114

Table 31: Correlation coefficients and loadings for immediate meaning tests 114

Table 32: Correlation coefficients and loadings for delayed meaning tests 114

Table 33: Effect of effort at three levels of success for form-recognition 116

Table 34: Effect of effort at three levels of success for immediate meaning tests 117

Table 35: Effect of effort at three levels of success for delayed meaning tests 118

Table 36: Frequency and concreteness indices for the English translations for the target words 139

LIST OF FIGURES

Figure 1: A conceptual illustration of the repetition pattern for one item 48

Figure 2: A summary of the experimental procedure 53

Figure 3: An example of one experimental trial sequence 54

Figure 4: Form-recognition scores in the three ISI conditions 66

Figure 5: L2- L1 translation scores in the three ISI conditions 66

Figure 6: Form-meaning mapping scores in the three ISI conditions 67

Figure 7: Posttest results in the three ISIs for the two groups of participants 68

Figure 8: Effect of study time on scores in the two groups 69

Figure 9: Median study-phase response latencies across the six repetitions 86

Figure 10: Response latencies in the short and long study duration conditions 87

Figure 11: Study-phase latencies in successful and unsuccessful retrieval attempts 88

Figure 12: Growth in the latencies in the three conditions across repetitions 89

Figure 13: Response latencies across five successful retrieval attempts 93

Figure 14: Response latencies as a function of condition and success of retrieval 95

Figure 15: Successful retrievals at each repetition in the three conditions 98

Figure 16: Growth in retrieval successes in the two study time conditions 99

Figure 17: Conceptual structure for the moderated parallel mediation analysis 107

Figure 18: Conceptual structure for the moderated mediation analysis 115

CHAPTER 1

SECOND LANGUAGE ACQUISITION, OVERT RETRIEVAL PRACTICE, AND THE SPACING/LAG EFFECTS

Learning large numbers of words is an important part of becoming proficient in a second language. Therefore, an important question for second language pedagogy is how to go about the task of learning/teaching vocabulary in a way that is both successful and efficient. Second language research has addressed this question by testing different methods of learning vocabulary. One method that has been widely found to increase retention of studied material in the field of psychology is to space repeated study of target material rather than use massed repeated study (Crowder, 1976; Dellarosa & Bourne, 1985; Dempster, 1988, 1989; Hintzman, 1974; Pavlik & Anderson, 2005; Rohrer & Pashler, 2007). This finding, widely known as the spacing effect, has also been observed with learning of second language vocabulary (Bloom & Shuell, 1981; Nakata, 2015; Nakata & Webb, 2016). A closely related finding, termed the lag effect, is the finding that the wider practice is spaced the better the learning outcomes. The spacing effect is one of the most robust and ubiquitous findings in memory research. The positive effects of spacing are usually very large: it is often found that, holding total exposure time constant, two exposures to a target item that are massed (consecutive) are hardly more effective than a single exposure while two spaced exposures are often about twice as effective as one. Spacing study offers important benefits also because it can help save time: no additional study time is required to observe the considerable learning benefits – in fact, less time may be required to attain more learning (Maddox & Balota, 2015). Because of its considerable benefits and practicality, the spacing effect potentially holds great promise for any learning situation. However, as noted by many, the full extent of its potential benefits is not being exploited in educational settings (Cepeda et al., 2009; Dempster, 1988; Gerbier & Toppino, 2015; Kang, 2016; Maddox, 2016). Further, despite the generality and consistency of the observed benefits of spaced practice obtained across vastly diverse populations and target tasks in the field of psychology, investigations of spaced practice in the context of second language learning have produced mixed results, with some studies finding that spacing repeated study more widely has no effect or even has a detrimental effect on learning (Collins, Halter, Lightbown, & Spada, 1999; Elgort & Warren, 2014; Rogers & Cheung, 2018; Serrano, 2011; Serrano & Munoz, 2007; Suzuki & DeKeyser, 2017; White & Turner, 2005). In order to understand when and how spacing repeated study of L2 material more widely may be beneficial for second language learning contexts and in order to be able to give useful practical recommendations regarding how to make the best use of this potentially very powerful learning tool in second language pedagogy, it is important to understand the underlying mechanisms that may drive any effects of spacing in specific learning situations. It is further important to understand how the operation of these mechanisms may be affected by variables that are relevant for any specific learning contexts. Prior SLA research has tested the effects of spacing repeated study on acquisition of various aspects of a second language and provides important insights into the usefulness of this learning method for SLA contexts. However, prior SLA research has not produced much direct investigation into the process as well as the product of learning from repeated exposures under different levels of spacing. The present study contributes to filling this gap. In the present study, I investigate the contribution of a proposed underlying mechanism of the spacing effect to novel L2 vocabulary learning from overt retrieval practice in a paired-associate learning (PAL) format.

Overt retrieval practice is another popular method that has been widely shown to produce powerful beneficial effects on learning. Information that is retrieved from memory becomes more recallable in the future. This finding is known as the retrieval effect (Carrier & Pashler, 1992; Cull, Shaughnessy, & Zechmeister, 1996). Just as is the case with the spacing effect, retrieval practice produces very large learning benefits and is a very robust and ubiquitous finding. Just as is the case with the spacing effect, it is not being taken full advantage of in education (McDaniel & Fisher, 1991; Roediger & Karpicke, 2006). Optimizing retrieval practice with L2 vocabulary is an important goal in L2 pedagogy. One way to make retrieval practice more effective is to space retrieval attempts more widely (Maddox & Balota, 2015; Maddox, Balota, Coane, & Duchek, 2011). The underlying mechanism here is proposed by some accounts to be a combination of retrieval effort and success (Bjork, 1994; Maddox & Balota, 2015), which is a dual mechanism that is also believed to more generally underly the effects of spacing any type of practice more widely (Benjamin & Tullis, 2010).

The present dissertation

In the present dissertation, I investigate the contribution of the two-process mechanism of effortful successful retrieval during study to the spacing/lag effect in L2 vocabulary learning. Such a dual mechanism is proposed to underlie the spacing/lag effect within the reminding framework (Benjamin & Tullis, 2010). I further investigate how the operation of the two mechanisms of study-phase retrieval effort and success, as well as ultimate learning outcomes, may be affected by a variable that is relevant for second language learning contexts, which is the amount of time a learner is allowed, per encounter (and in total, while holding the number of encounters constant), for studying a foreign word with its translation. This latter variable is referred to, throughout this text, as study time or presentation duration.

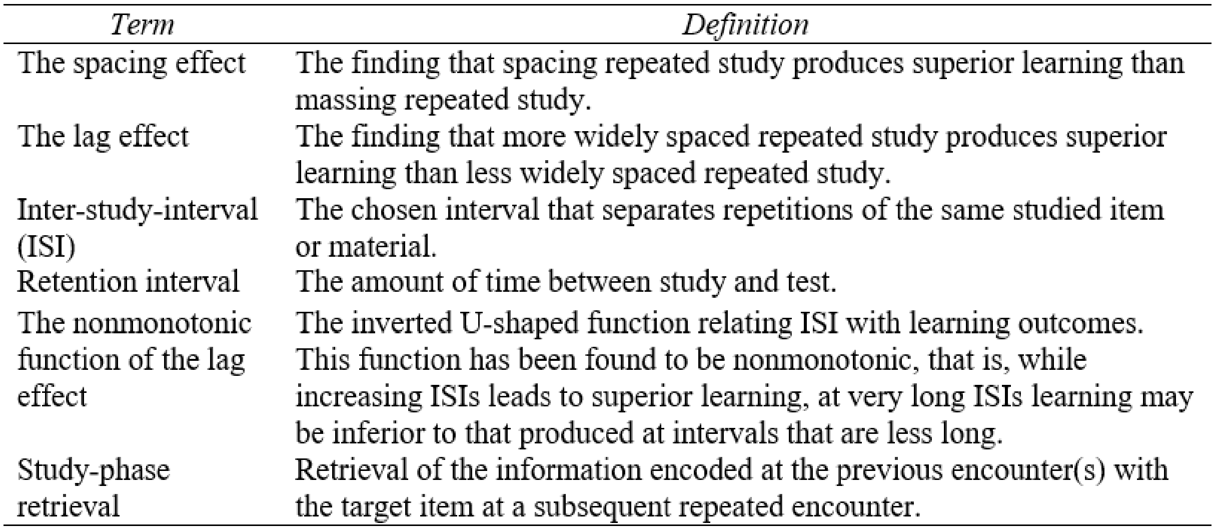
Using a fully counterbalanced within-participant within-item design, I investigate learning of novel foreign vocabulary in a PAL format (Barcroft, 2007; Nakata, 2011) within one session under three levels of inter-study interval (ISI): (a) 0-1 intervening trials, (b) 17-38 intervening trials or 12-22 trials and a six-minute break (c) 71-119 intervening trials and the six-minute break. I further investigate any mediating effects of successful effortful overt retrieval of the paired L1 translation associate (Maddox & Balota, 2015; Maddox et al., 2011; Nakata, 2015) by using response accuracy and latency as proxies for retrieval success and effort, respectively (Maddox & Balota, 2015, Maddox, Pyc, Kauffman, Gatewood, & Schonhoff, 2018) as well as the role of feedback study time in moderating these effects (Verkoeijen & Bouwmeester, 2008). I use two levels of feedback presentation duration: (a) 3 seconds and (b) 9 seconds. This refers to the length of time a foreign word and its L1 translation stay on the screen for the learners to study following each of its retrieval attempts. The total study time for the words is 18 versus 54 seconds over six exposures. The amount of time a learner is allowed to study a word with its translation is an important variable for second language vocabulary learning success that has not received much attention in SLA research. While it has been shown that the time learners choose to spend on studying or attentionally processing a target item has an important positive effect on learning of the item (Godfroid et al., 2018; Godfroid et al., 2013; Koval, 2019; Rundus, 1971), it is not obvious that the same effect should be observed when study time is externally imposed on the learner by a word-learning software or an instructor. In the present study, I investigate whether the benefits of longer study time will hold when the length of study is externally determined.

Further, the length of time a learner is given to study a target L2-L1 translation pair may have important effects on the study-phase processes of retrieval success and effort. Longer study time at each repetition is likely to result in stronger encodings (Verkoeijen & Bouwmeester, 2008), which, in turn, might increase the likelihood of retrieval success on subsequent repetitions but also decrease the amount of effort needed for such retrieval. In this way, in addition to having potential learning benefits due to increased exposure to the target translation pair, study time could affect the operation of the proposed underlying mechanisms. The present dissertation aims to answer the following general research questions: (a) Does the dual mechanism of successful effortful retrieval underlie the benefits/detrimental effects of spacing on L2 vocabulary learning in a PAL format?, (b) Does exposure duration moderate these effects?

Overview of the dissertation. The present dissertation consists of three chapters. In the first chapter, I introduce the motivation for the present dissertation and its main goals. In the second chapter, I discuss extant literature and present the methodology and results of the present experiment. In the third chapter, I present a discussion of the present results and well as their pedagogical implications, followed by a discussion of limitations of the present experiment and suggestions for future research.

Definition of key terms. Table 1 presents a list of key terms with their definitions.

Table 1: *Definitions for key terminology*



Term	*Definition*
The spacing effect	The finding that spacing repeated study produces superior learning than massing repeated study.
The lag effect	The finding that more widely spaced repeated study produces superior learning than less widely spaced repeated study.
Inter-study-interval (ISI)	The chosen interval that separates repetitions of the same studied item or material.
Retention interval	The amount of time between study and test.
The nonmonotonic function of the lag effect	The inverted U-shaped function relating ISI with learning outcomes. This function has been found to be nonmonotonic, that is, while increasing ISIs leads to superior learning, at very long ISIs learning may be inferior to that produced at intervals that are less long.
Study-phase retrieval	Retrieval of the information encoded at the previous encounter(s) with the target item at a subsequent repeated encounter.

CHAPTER 2

THE SPACING AND LAG EFFECTS AND EFFORTS TO UNDERSTAND THEIR UNDERLYING MECHANISMS

Research interest in the spacing effect is known to have been sparked by Ebbinghaus' (1885/1964) influential book on memory. Research of this memory phenomenon has been quite prolific since that time. The benefits of spacing practice have been consistently obtained under a wide range of learning conditions and target tasks (Crowder 1976; Dempster 1996; Donovan & Radosevich, 1999; Hintzman, 1976); with younger and older individuals (Balota, Duchek, & Paullin,1989) and in healthy humans as well as in people with memory impairments (Green, Weston, Wiseheart, & Rosenbaum, 2014; Hillary et al., 2003). Memorial benefits of spacing have also been found in other species, such as monkeys, rodents, and even honeybees and drosophilae (Commins, Cunningham, Harvey, &Walsh, 2003; Deisig, Sandoz, Giurfa, & Lachnit, 2007; Yin, Del Vecchio, Zhou, & Tully, 1995). Thus, the spacing effect appears to be a robust and quite universal finding. Further, its beneficial effects are usually found to be large, suggesting that spacing study is potentially a very powerful learning tool that may be used in a wide range of learning situations.

Studies investigating the spacing effect usually compare learning under two conditions: a massed condition, where repetitions of the studied material are consecutive, and a spaced condition, where repetitions are separated by time or study of other material. Psychology studies of the effects of spacing repetitions also usually include once-presented words (Braun & Rubin, 1998). These serve as filler material to achieve the desired order and spacing of the target items as well as a baseline for investigating the effects of repetition. In its strictest sense, massed practice refers to situations where repetitions of the same item are separated by zero intervening items or time that is no longer than one second (Carpenter, Cepeda, Rohrer, Kang, & Pashler, 2012; Kahana & Howard, 2005), while spaced repetitions are those that are separated by a longer period of time or at least one intervening item. A closely related phenomenon, known as the lag effect, is the finding that longer ISIs lead to better long-term retention than shorter ISIs (D'Agostino & DeRemer, 1973; Toppino, & Gracen, 1985). Studies investigating the lag effect usually include more than one level of lag – that is, repetitions are separated by different intervals of time or numbers of intervening items in different lag conditions. In studies investigating learning from more than two repetitions, the spacing between each two consecutive repetitions may be constant (or equal) or it may be progressively longer (what is known as an expanding schedule) or shorter (what is referred to as a shrinking schedule). Further, the increase or decrease in the amount of spacing across repetitions may be systematic (such as 0-2-4-6 intervening items) or unsystematic (such as 0-1-5-6 intervening items). In studies investigating nonuniform lag schedules, the average lag is held constant across the different tested lag schedules for more valid conclusions regarding the effects of nonconstant spacing schedules that are not confounded with different overall amount of time between repetitions. Further, the number of repetitions may be constant or not, or it may depend on participants' performance levels. In what is known as a drop-out schedule, target items are tested during the acquisition phase (usually through overt response) until a criterion level of knowledge is reached, at which time the items in question do not appear for further study. This latter method may be useful in investigations of forgetting, where each item needs to be at the same level of mastery at the end of the acquisition phase, thus equating intercepts of the forgetting curves for the different items in the different learning conditions (e.g., Pyc & Rawson, 2009). Some studies have varied the number of repetitions a priori to investigate the effects of repetition at different levels of ISI (e.g., Maddox & Balota, 2015). This allows to test whether fewer or more repetitions are needed with a given ISI schedule.

Theories of the spacing effect

Despite the fact that research interest in the spacing and lag effects dates back over a century and despite the large number of theories that have been proposed in efforts to explain it (Benjamin & Tullis, 2010; Bjork & Allen, 1970; Challis, 1993; Dellarosa & Bourne, 1985; Estes, 1955; Glenberg, 1979; Greene, 1989; Jacoby, 1978; Küpper-Tetzel, & Erdfelder, 2012; Landauer, 1969; Madigan, 1969; Melton,1970; Pavlik & Anderson, 2005; Raaijmakers, 2003; Rundus, 1971; Thios & D'Agostino, 1976; Zimmerman, 1975), its underlying mechanisms are still poorly understood (Kılıç, Hoyer, & Howard, 2013; Maddox et al., 2018). Further, it is widely recognized today that a different mechanism, or combination of mechanisms, may underlie the effects of spacing depending on a specific learning situation or target task (Gerbier & Toppino, 2015; Glenberg & Smith, 1981; Greene, 1989; Kornell & Bjork, 2008; Russo & Mammarella, 2002). One proposed mechanism that is intuitively relevant for second language learning is that proposed by the deficient processing theory of the spacing effect (Bjork, 1999; Callan & Schweighofer, 2010; Challis, 1993; Cuddy & Jacoby, 1982; Hintzman, 1976; Jacoby, 1978, Pavlik & Anderson, 2005; Rose & Rowe, 1976; Rundus, 1971; Zechmeister & Shaughnessy, 1980). According to this theory, repetitions of the same stimulus that occur in close succession receive less attentional processing than repetitions that occur more widely apart. Such an attentional account assumes that more attentional processing leads to better learning outcomes, which is in line with proposals in the field of SLA (Gass, 1988; Robinson, 2003; Schmidt, 1990, 2001), in general, and findings from L2 vocabulary studies (Godfroid et al., 2018; Godfroid, et al., 2013), in particular. In fact, in Koval (2019), I found that more attentional processing that is given to novel L2 words that occur with longer intervals between repetitions mediates the large beneficial effects of spacing obtained in my study, suggesting that the mechanism proposed to underlie the beneficial effects of spacing by the deficient processing theory contributes in important ways to the effects of spacing on learning L2 vocabulary.

According to theory, deficient processing may be due to voluntary or involuntary mechanisms. Thus, less than optimal processing of massed repetitions may be the result of a conscious choice to give less attention to an immediate repetition of the same stimulus due to a heightened sense of familiarity (Greene, 1989; Kornell & Bjork, 2008; Rundus, 1971; Shaughnessy, Zimmerman, & Underwood, 1972; Zechmeister & Shaughnessy, 1980; Zimmerman, 1975). Such a voluntary, consciously controlled mechanism is particularly relevant for intentional learning situations, such as when one is trying to learn a list of L2 words. Thus, when a word is repeated immediately, one may overestimate one's knowledge of the word and strategically choose to allocate less study time to it. When, on the other hand, a word is repeated after a substantial amount of time has gone by and, consequently, the memory trace of the previous encounter has faded quite a bit more relative to what occurs within the short time between massed repetitions, the word may strike the learner as less familiar, in which case more rehearsal will seem warranted. An involuntary deficient processing mechanism, on the other hand, operates automatically, such as through the process of habituation, priming, or neural repetition suppression (Callan & Schweighofer, 2010; Challis, 1993; Mammarella, Avons, & Russo, 2004; Russo & Mammarella, 2002; Russo, Parkin, Taylor, & Wilks, 1998; Van Strien, Verkoeijen, Van der Meer, & Franken, 2007; Xue et al., 2011). Thus, for example, recognition of an immediate repetition usually requires a much less extensive analysis of the target stimulus than its recognition upon its first presentation or when it is repeated after a longer time interval and some forgetting of the initial presentation has occurred. Processing is further often said to be deficient in terms of the amount of effort involved in retrieval of information (Bjork, 1994,1999). More effortful, or difficult, retrieval is believed to be desirable for stronger memory traces (Benjamin, Bjork, & Schwartz, 1998; Benjamin & Tullis, 2010; Bjork, 1994, 1999; Gardiner, Craik, & Bleasdale, 1973; Jacoby, 1978; Logan & Balota, 2008; Pavlik & Anderson, 2005; Roediger & Karpicke, 2006; Schmidt & Bjork, 1992). Repeated retrieval practice that is massed is often assumed to require less effort than repeated retrieval practice that is spaced (Benjamin & Tullis, 2010; Bjork, 2013; Pyc & Rawson, 2009) or to involve less complete retrieval processes because the to-be-retrieved information still resides in working memory (Glover, 1989).

An important characteristic of the lag function (the function relating various degrees of ISI and learning success) is that it is nonmonotonic, or an inverted-U in shape (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Cepeda et al., 2009; Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008; Küpper-Tetzel, & Erdfelder, 2012; Rohrer & Pashler, 2007). This means that with shorter ISIs, increasing the ISI leads to more learning; however, as lags get increasingly longer, there comes a point beyond which increasing lag any further may actually have detrimental effects on learning (Benjamin & Tullis, 2010; Cepeda, et al., 2006; Maddox, 2016; Peterson, Wampler, Kirkpatrick, & Saltzman, 1963; Young, 1971). In other words, there is a limit to how widely we can space repeated study before this begins to actually have a detrimental effect on learning outcomes. The finding that learning gains do not increase monotonically with longer lags but increase only to a point beyond which learning actually begins to decrease with increasing ISIs cannot be explained by the deficient processing theory, as, while the increase in the amount of attention given to progressively wider spaced repetitions may well level off at some point, it is unlikely to begin to decrease at a longer ISI. As a response to findings of such limitations to single process theories, many current theories assume the operation of multiple processes that together contribute to the effects of spacing (Delaney et al., 2010; Greene, 1989; Maddox, 2016). In fact, it is argued that no single-process mechanism can accommodate the broad range of findings from research into the spacing effect and its boundary conditions (see, e.g., Benjamin & Tullis, 2010; Delaney, Verkoeijen, & Spirgel, 2010; Gerbier & Toppino, 2015; Greene, 1989; Maddox, 2016; Verkoeijen, Rikers, & Schmidt, 2004). A leading explanation that can accommodate both the finding that attentional engagement mediates the benefits of spacing on the one hand and the fact that the lag function is nonmonotonic, on the other, is the reminding account (Benjamin & Tullis, 2010). This account supplements the operation of a deficient processing mechanism with the central assumption of the study-phase retrieval theory (Braun & Rubin, 1998; Delaney et al., 2010; Greene, 1989; Raaijmakers, 2003; Thios & D'Agostino, 1976; Toppino & Bloom, 2002). This is the assumption that, for spacing to have its benefits, a repeated encounter must involve retrieval of its previous presentation from long term memory (Wahlheim, Maddox, & Jacoby, 2014). Other evidence of the importance of such dependency among memory traces comes from the finding of super-additive effects in learning from repetition. Super-additivity refers to the fact that the probability of recalling an item that was studied twice is found to be higher than the probability of recalling any of two items studied once (the additive assumption) (Begg & Greene, 1988; Ross & Landauer, 1978; Watkins & Kerkar, 1985; Waugh, 1963). Such a finding that memory for an item studied twice usually exceeds what would be expected from two independent learning events indicates that effects of repetition on learning are more than just the sum of learning events.

Theories that can accommodate the curvilinearity in the lag function explain the shape of the function in terms of the importance of preserving memory trace dependency between repetitions (the study-phase retrieval assumption discussed above). Thus, at relatively shorter lags such a dependency is preserved and repetitions are processed as repetitions rather than as independent events while at longer lags this dependency may be broken, which has a negative effect on learning outcomes. A number of other findings in the field of psychology that are potentially relevant for second language learning can be accommodated by a theory that assumes the importance of memory trace survival between repetitions, or successful study-phase retrieval. One such finding is that the optimal ISI (the inflection point in the lag function at which learning is best and beyond which learning begins to decrease with increasing ISIs) under intentional learning is farther out (at a higher level of ISI) than that under incidental learning (Verkoeijen, Rikers, & Schmidt, 2005). This can be explained in terms of the stronger memory traces laid down under intentional learning conditions, which are traces that are more likely to survive over longer ISIs. Another important finding is that when repeated exposures occur within contexts that are intentionally made different through experimental manipulation, spacing repeated exposures more widely may have a detrimental effect on learning outcomes (Verkoeijen, et al., 2004). This finding can also be accommodated by a theory that assumes an important role for successful retrieval of the previous study event, because when an item repeats in a context that is different from its previous encounter it is less likely to be recognized as repeated, in which case the dependency between the memory traces may not be preserved. Further, study time has been found to positively affect learning from spaced repetitions (Verkoeijen & Bouwmeester, 2008) while task complexity and the difficulty of the intervening task coupled with lower working-memory capacity have been shown to negatively affect learning from spaced repetitions (Bui, Maddox, & Balota, 2013; Donovan & Radosevich, 1999). Thus, the findings that positive effects of spaced study may be tempered or even reversed under certain levels of the relevant variables can also be explained through this affecting the probability of study-phase retrieval success.

The two-process reminding account

The reminding account (Benjamin & Ross, 2010; Benjamin & Tullis, 2010; Hintzman, 2004; 2010; Tullis, Benjamin, & Ross, 2014) is currently a leading explanation for the lag and spacing effects. It is a dual mechanism account that combines beneficial effects of desirable difficulty (Bjork, 1994, 1999) with an important role for study-phase retrieval, or reminding (Hintzman, 2004, 2010; Thios & D'Agostino, 1976). Both desirable difficulty and reminding are believed to benefit memory independently of any effects of spacing (Bjork, 1994; McKinley, Ross, & Benjamin, 2019). Bjork (1994) has argued that retrieval is most beneficial when the to-be-retrieved item is difficult but still not impossible to remember. According to the reminding explanation of the spacing effect, learning from repetition is optimal when the second encounter with an item triggers retrieval of (or reminds of) its first occurrence and, at the same time, such retrieval requires more effortful processing (or the information is retrieved from long-term rather than short-term memory). With increasing ISIs, retrieval of a previous encounter requires more effort, which is beneficial for learning. At the same time, however, retrieval is only likely to be successful within a limited range of ISIs, beyond which such retrieval may fail, resulting in detrimental effects on learning. In this way, the dual process assumed by the reminding account can accommodate the above discussed findings of nonmonotonicity in the lag function as well as the other previously discussed findings that are potentially of relevance for second language acquisition. Importantly, the reminding account may be able to explain the mixed findings obtained in the field of SLA regarding the effects of spacing repeated study of SLA material. A failure to retrieve the previous encounter with a repeated item, or to process a repeated encounter as repeated, may be the reason, as has been speculated though not directly tested in a number of SLA studies (see, e.g., Elgort & Warren, 2014; Serrano, 2011), for failure to observe benefits of spacing in some SLA research.

Investigating the role of attention and effort in learning from repetition

Attention is known to be important for learning a second language (Gass, 1988; Robinson, 2003; Schmidt, 1990, 2010). Amount of attention or study given to a target L2 word has been shown to be positively related to memory for the words (Godfroid et al., 2018; Godfroid, et al., 2013; Koval, 2019). In both psychology and SLA, studies investigating repeated study of target items show that the more time a learner spends studying a given word per repetition, the better the learning outcomes (Godfroid et al., 2018; Godfroid, et al., 2013; Koval, 2019; Rundus, 1971). Such studies further showed that when learning targets are encountered or studied multiple times, reading or study time decreases across repetitions, though the steepness of the slope may depend on the temporal distribution of repetitions (Koval, 2019; Rundus, 1971; Shaughnessy et al., 1972). In Koval (2019), I showed that the amount of study given to L2 words studied in sentence contexts, which was greater with spaced repetitions than with massed repetitions, mediated the learning benefits obtained by spacing the repetitions more widely.

Studies testing deficient processing of massed repetitions as an explanation for the beneficial effects of spacing have employed different methods to measure effort and amount of attentional processing that learners choose to allocate to target items. In some studies, participants have studied words presented one per slide and pressed a button to indicate that they wished to move on to the next slide with a new word. The time between the onset of each slide and the button press was recorded and used as an index of study time for the word in question (Rundus, 1971; Shaughnessy et al., 1972; Zimmerman, 1975). In some such studies, participants were asked to rehearse aloud during study and the time during which such overt rehearsal was produced was used as a more precise measure of processing time (Rundus, 1971, Experiment 3; Zimmerman, 1975). In Koval (2019), I recorded participants' eye movements as they read L1 sentences with embedded L2 words that participants studied for a subsequent test. For my main analysis, I used the measure of total reading time, which is an index of the amount of time a word was looked at within a given sentence in total, that is, during the first time the gaze landed on the word and each time the word was subsequently revisited, before the participant chose to move on to the next sentence. Based on this measure, I inferred the amount of attention the words received in the massed condition, where the same word repeated in consecutive sentences, and in the spaced condition, where the same word repeated in sentences separated by other sentences containing other target L2 words plus a distractor math task.

Studies investigating how learners choose to allocate study time have shown that learners tend to overestimate their knowledge of items that are repeatedly studied in close succession (massed practice) and consequently give less study time to these items (Benjamin et al., 1998; Kornell, & Bjork, 2007; Koval, 2019; Rundus, 1971; Shaughnessy et al., 1972;

Zechmeister & Shaughnessy, 1980; Zimmerman, 1975). Generally, learners are known to be quite ineffective at pacing their own study (Benjamin et al., 1998; Jacoby, Bjork, & Kelley, 1994; Kornell, & Bjork, 2007). Consequently, an interesting question that has important practical implications for the development of pedagogical tools, including computer programs that present L2 words for learning using the PAL method, is whether the amount of time a learner is given for study of an L2 word per encounter affects learning in the same way as does learner-regulated study (De Jonge, Tabbers, Pecher, & Zeelenberg, 2012). Intuitively, one would expect that the amount of time available for study of a novel word should be positively related to learning: the longer a learner spends on the task of learning a given word, the better they will remember it on a subsequent test (Ebbinghaus, 1885/1964), in line with what has been found with self-paced study. If learners tend to not be effective at pacing their study, can we improve learning by controlling the pace at which words are studied? Predetermining study time for a given item to be longer may help counteract poor study strategies and the ineffective pacing that learners tend to adopt. Studies that have measured the time participants choose to allocate to study of massed and spaced items have argued that the underlying reason for benefits of spaced practice is that learners choose to spend more time studying the items in the a spaced condition relative to a massed condition. If, in a purely quantitative way, longer study time underlies the beneficial effects of spacing, by holding study time constant at two levels across the ISI conditions, we may fail to observe any effects of ISI but instead observe a strong effect of presentation duration, or study time. Alternatively, it may be the case that the quality of processing may change beyond the point at which a learner would have chosen to move on to a different item if they were free to control their own pace. It is quite likely that the processes that are engaged during the initial stages of presentation of an L2-L1 translation pair, where the learner establishes or revises form-meaning mappings, differ qualitatively from those engaged once this process is complete and the learner simply repeats the information to themselves to maintain it in short-term memory. However, there may further be a qualitative difference between processing that is beyond such an initial recognition and encoding stage though it is still learner-regulated, where learners feel like they have not reached a kind of a saturation point at which they would wish to stop studying a given word and move on to the next item, and processing that occurs after such a saturation point, where rehearsal is externally imposed on the learner.

Psychology studies have examined the effects of other-imposed total time given for study on subsequent recall (Bugelski, 1962; De Jonge et al., 2012; Johnson, 1964; Murdock, 1960), as well as presentation duration per trial while holding total time allowed for study constant (Zeelenberg, de Jonge, Tabbers, & Pecher, 2015). As intuition would suggest, the amount of time a participant is given for study of target items was often shown to be positively related to later recall of the items (Bugelski, 1962; Johnson, 1964). This is in line with proposals that the time an item spends in primary, or short-term, memory during study is positively related to later recall (Atkinson, & Shiffrin, 1968; Braun & Rubin, 1998; Rundus, 1971; Rundus, & Atkinson, 1970, Waugh & Norman, 1965). There are, however, important findings to the contrary. Thus, for example, in their well-known Experiment 1, Craik & Watkins (1973) had participants listen to a list of L1 words for an immediate memory test, where they would have to report the last word that started with a given letter. This forced participants to maintain each word that starts with the letter in question (critical word) in memory until they encountered the next word that started with the same letter, at which point they switched to rehearsing this new word. The number of intervening noncritical words (which did not begin with the critical letter) was varied, resulting in different lengths of time during which a critical word had to be maintained in working memory. The results showed no benefit of longer intervals over short intervals on a surprise recall test given after a short break following the last list of study items. Such a finding goes against evidence that amount of rehearsal has benefits for learning (Atkinson, & Shiffrin, 1968; Rundus, 1971; Waugh & Norman, 1965). Following Craik and Lockhart's proposal (1972), Craik and Watkins suggested that the mode of rehearsal may be key: simply repeating a word to oneself to maintain it in primary memory (known as *maintenance* rehearsal) may not hold much benefit for longer-term retention. Thus, the amount of rehearsal, or the time an item spends in short-term memory, is argued to only have benefits for long-term retention when the item is being processed elaboratively (or associatively). Craik and Watkins' Experiment 2 further showed that an increased number of overt maintenance rehearsals did not improve long-term retention of target items. The authors conclude that maintenance of a studied item in short-term storage does not necessarily increase its strength in the long-term store. While maintenance rehearsal may have limited benefits for the final test of free-recalling which of the many well-known L1 words had been seen during an experiment, it is not obvious that the amount of time a learner is allowed to rehearse a novel L2 word form presented with its L1 translation will produce the same pattern of results. Thus, it is an interesting question whether increasing study time, or adding rehearsal time, for an L2-L1 translation pair at each repetition will benefit learning of the L2 word.

Longer study time at each repetition may additionally have an effect on the effort and success of retrieval during subsequent repetitions within the study phase and thus may affect the operation of the investigated underlying mechanisms of spacing practice (Verkoeijen &

Bouwmeester, 2008). Verkoeijen and Bouwmeester manipulated presentation rate during study (1 second vs. 4 seconds per word). Based on posttest results, they identified a high performance group and a low performance group among their participants. They found that while the former group benefitted from spaced practice regardless of the presentation rate, the latter group benefitted from spacing only when the presentation rate was longer. They suggest that longer presentation duration serves to establish stronger memory traces at each repetition which may be more likely to survive longer lags between repetitions.

Effort has been shown to benefit learning in diverse experimental paradigms in psychology. For example, Auble and Franks (1978) showed that providing more time for effort toward sentence comprehension resulted in better subsequent recall performance. More work and effort that is required by a task has widely been shown to be beneficial for learning outcomes (e.g., Benjamin et al., 1998; Gardiner et al., 1973; Soderstrom, Kerr, & Bjork, 2016; Whitten & Bjork, 1977). A number of studies have operationally defined effort as response latencies in the performance of various tasks (Braun & Rubin, 1998; Glover, 1989; Karpicke & Roediger, 2007; Logan & Balota, 2008; Maddox & Balota, 2015; Maddox et al., 2018; Pyc & Rawson, 2009). In investigating effects of spacing and lag on response latencies and success as well as on subsequent learning gains, Braun and Rubin (1998) found that effort in covert retrieval of a previous presentation of L1 words that were related in form increased with lag however no lag effect was observed in learning gains beyond a spacing effect. The opposite pattern was observed in Maddox et al. (2018). Using L1 word recognition latencies as a proxy for retrieval effort, the authors found a lag effect in posttest scores but no difference in study-phase recognition latencies beyond the effect of spacing versus massing of repetitions, contrary to the predictions of the reminding account. In their 2015 study, Maddox and Balota had participants study arbitrary L1 word pairs. They recorded latencies for overt retrieval of paired associates (cued recall) across a number of repeated retrieval attempts in younger and older adults. The results of their experiments were overall consistent with the reminding account. However, the task of learning novel L2 forms with their meanings may involve a different dynamic of underlying processes than the task of recognizing known L1 words or retrieving their arbitrary L1 word associates as well as the process of studying arbitrary pairings of known L1 words. The learning outcomes measured in the field of psychology are also often different: while in L2 vocabulary learning we are concerned with learners' acquisition of novel word forms and the development of form-meaning mappings, in psychology research the target knowledge may be associations of arbitrary well-known L1 words, ability to free recall as many as possible, or even memory of their relative order during acquisition. Thus, results from psychology studies may often have limited relevance for learning of an L2 (Nelson & Dunlosky, 1994).

Often, in the spacing effect research, target words or other items are studied only twice, although there are exceptions. Maddox and Balota (2015), for example, investigated paired-associate learning of known L1 words over a number of repetitions. In this study, however, the retrieval attempts were not followed by feedback. Feedback is often not provided in psychology research due to the specific research questions that are often different from those in second language learning. Further, because no feedback is provided, here only items that are correctly retrieved during the study phase are usually analyzed in terms of effort and learning outcomes (Braun & Rubin, 1998; Maddox & Balota, 2015; Maddox et al., 2018; Pyc & Rawson, 2009). In an L2 vocabulary learning context, however, because feedback is usually provided, it makes sense to analyze both successful and unsuccessful retrieval attempts during the study phase. This is because, while in a study such as Maddox and Balota (2015), items that are not successfully retrieved in early repetitions during study phase are very unlikely to be successfully retrieved in later repetitions and are mostly simply forgotten, this pattern is reversed with the type of practice that is done in L2 vocabulary learning and where feedback is provided: here, retrieval success will likely grow across repetitions as learners learn from the feedback that is provided following each retrieval attempt.

Investigating the role of reminding in learning from repeated study

Reminding, or the retrieval of the previous encounter(s) with the to be learned material, has been shown to be important for retention of studied material (Batchelder, & Riefer, 1980; Bellezza, Winkler, & Andrasik, 1975; Bruce & Weaver, 1973; Glanzer, 1969; Glover, 1989; Jacoby, 1974; McKinley et al., 2019; Robbins & Bray, 1974; Wahlheim et al., 2014). Such reminding may be triggered by a repeated encounter with the same material or an encounter with related material (Benjamin & Tullis, 2010; Braun & Rubin, 1998; McKinley et al., 2019). The effects of reminding on memory have been observed with various tasks employed in psychology research, such as classification (category) learning (Medin & Schaffer, 1978; Ross, Perkins, & Tenpenny, 1990), ambiguity resolution (Ross & Bradshaw, 1994; Tullis, Braverman, Ross, & Benjamin, 2014), and problem solving (Ross, 1984), and with various outcome measures, such as cued recall (Jacoby & Wahlheim, 2013), free recall (Tullis et al., 2014), absolute and relative temporal (recency or order) judgments (Hintzman, 2010; Jacoby & Wahlheim, 2013), frequency judgements (Hintzman, 2004), and list discrimination (Jacoby & Wahlheim, 2013).

Current theories of the spacing effect include a key role for reminding, or retrieval of an item's earlier presentation upon repeated encounters, during the study phase (referred to as study-phase retrieval) for observing the beneficial effects of spacing repeated study. Positive effects of successful study-phase retrieval on learning from spaced study have been found in various tasks employed by psychology research to investigate the spacing or lag effects (Appleton-Knapp, Bjork, & Wickens, 2005; Benjamin & Tullis, 2010; Braun & Rubin, 1998; Greene, 1989; Hintzman, 2004, 2010; Hintzman, Summers, & Block, 1975; Pavlik & Anderson, 2005; Raaijmakers, 2003; Siegel & Kahana, 2014; Thios and D'Agostino, 1976). Thus, in addition to enhancing learning from repetition, reminding may be crucial for observing beneficial effects of spacing (Thios and D'Agostino, 1976). A higher chance of study-phase retrieval success is believed to be the reason underlying the findings of benefits of expanding spacing schedules. Thus, for example, Maddox et al. (2011) found that positive effects of an expanding schedule were conditional on initial repetitions being close enough to the original encoding to produce successful retrieval. This is explained based on the logic that if an item can be retrieved more easily from working memory upon its second presentation, scheduling initial repetitions closely together might ensure a stronger encoding that is more likely to survive increasingly longer subsequent lags. However, again, Maddox et al.'s design did not include feedback. A different pattern may be observed when each retrieval attempt is followed by the presentation of the target material, as here initial retrieval success may not be as crucial.

While in much psychology research study-phase retrieval is inferred based on the experimental design, some studies have attempted more direct investigation of the reminding process. This has been accomplished with the help of a number of techniques, such as the continuous recognition or repetition detection paradigm (Bellezza et al., 1975; Braun & Rubin, 1998; Kiliç et al., 2013; Maddox et al., 2018; Wahlheim et al., 2014). Here, participants are presented with stimuli, such as advertisements (Appleton-Knapp et al., 2005), L1 words (Maddox et al., 2018), or novel letter strings such as CCC and CVC strings (Bellezza et al., 1975), that repeat at different lags and whose repeated presentations are interleaved with the presentation of other advertisements, L1 words, novel letter strings, etc. Participants are to perform a repetition detection task (or old/new judgment), that is, they are to judge whether a given item has or has not occurred previously during the study phase. Bellezza et al. (1975) were among the first to demonstrate that items that are recognized as repeated upon their second presentation have a memorial advantage in the posttest performance.

Success/failure of study-phase retrieval has also been investigated with the help of what are known as indirect or implicit memory tests (Richardson-Klavehn & Bjork, 1988). In an indirect memory test, participants do not engage in an active search of their memory. Instead, retrieval of previously presented information is inferred based on changes in task performance, such as faster task performance (Koval, 2019; McKinley et al., 2019).

Finally, the effects of study-phase retrieval success have also been investigated by asking participants to overtly retrieve studied information, such as the second member of a pair of words studied in a PAL format (Maddox & Balota, 2015; Maddox et al., 2011). Maddox and Balota (2015) used successful overt retrieval of the paired associate as an index of successful study-phase retrieval (or reminding). Additionally, as done in previous research (Glover, 1989; Karpicke & Roediger, 2007; Logan & Balota, 2008; Maddox et al., 2011; Maddox et al., 2018; Pyc & Rawson, 2009), they used study-phase response latencies as a proxy for retrieval difficulty, which enabled them to test successful effortful overt retrieval in L1 paired-associate learning as a proposed mechanism for the effects of spacing.

Success of study-phase retrieval may depend on certain variables that may affect the probability such retrieval. One such variable may be how similar the context at repetition is to that at a prior encounter (Appleton-Knapp et al., 2005; Verkoeijen et al., 2004). Crucially, the probability of study-phase retrieval at a repeated encounter also depends on the strength of the memory trace that was laid down at a previous encounter. Verkoeijen et al. (2005) showed that when items are studied intentionally they show larger spacing effects and a longer optimal ISI. This may be attributed to stronger memory traces laid down during intentional study. Verkoeijen and Bouwmeester (2008) manipulated presentation rate during study (1 second vs. 4 seconds per word) and found that participants who had lower performance on the posttest benefitted from spaced practice only when presentation duration was longer. Verkoeijen and Bouwmeester discuss these results in terms of differential success of study-phase retrieval and the role of presentation rate for establishing stronger encodings that make such success more likely. However, the authors acknowledge that a limitation of their design is that they did not include a direct measure of study-phase retrieval but only inferred it based on the logic that participants who recalled more items at test likely had a higher rate of successful retrieval during study.

Working under the assumption that study-phase retrieval plays an important role in learning from repetition, Bui et al. (2013) asked the question of whether individual differences in the ability to retrieve a previous exposure affected learning from spaced repetition. Holding the ISI constant at 30 seconds, the researchers manipulated the difficulty of the intervening 30-second task, reasoning that this should modulate participants' ability to retrieve the earlier information due to differential degrees of interference. These authors did not interleave studied words but, instead, used an unrelated intervening task between repeated study of target words. They found that individuals with lower working memory capacity showed greater learning when the intervening task difficulty was low while individuals with higher memory capacity benefitted from a difficult intervening task. These results, too, are interpreted in terms of difficult reminding, or successful effortful retrieval. Here, again, study-phase retrieval was inferred rather than directly tested.

The nature of study-phase retrieval – that is, what exactly must be retrieved – is yet to be fully specified. Some efforts have been made in this direction, however. Delaney, Godbole, Holden and Chang (2018) investigated the nature of study-phase retrieval. Specifically, they asked whether the reminding mechanism relies on recollection, which is a process that involves retrieval of an earlier presentation, or on simple recognition, which does not involve an active memory search process but relies only on a judgment of familiarity (Oberauer, 2005; Yonelinas & Jacoby, 2012). The authors addressed this question by testing potential moderating effects of working memory span on the effect of spacing. If successful study-phase retrieval relied on explicit retrieval of episodic information, which depends on an individual's operational span (McCabe, Roediger, McDaniel, Balota, & Hambrick, 2010), a lag by span interaction was expected, where longer lags benefit learning in individuals with high working memory capacity but not in individuals with low working memory capacity. The authors found that spacing and working memory had an additive, rather than a multiplicative, effect on learning, suggesting no involvement of capacity-dependent mechanisms. The authors conclude that study-phase retrieval relies on a process of recognition rather than recollection. These results contradict the finding by Bui et al. (2013), who found that working memory capacity did play a role when repeated study was separated by a more difficult task.

The nature of study-phase retrieval and the ways in which its operation may be affected by variables that are relevant for specific learning situations are still far from being fully understood. Further, its operation during study of a second language has not been investigated directly. The present study is a first step towards understanding the complex nature of the relationships between retrieval effort and success with regard to novel L2 vocabulary learning by investigating the role of overt form-meaning mapping retrieval effort and success over six repetitions that occur at three different levels of ISI in the presence of feedback that follows each retrieval attempt, as well as the ways in which the time a learner is given for study of the target L2-L1 pairs may affect these relationships.

Overt retrieval practice and its effects on memory. The present study investigates the mechanism of retrieval effort and success as underlying any effects of lag in overt retrieval practice. Overt retrieval practice has been widely shown to enhance learning of target material. This known as the retrieval effect (Carrier & Pashler, 1992; Cull et al., 1996). The act of retrieval is known to be a “memory modifier” (Bjork, 1975), which refers to the fact that the memory trace of the information that is retrieved is altered such that it becomes more strongly represented and better connected with more robust, more elaborate, and more numerous retrieval routes, and is, consequently, more accessible for future recall (Birnbaum & Eichner, 1971; Bjork, 1975; Izawa, 1971, 1985; Karpicke, & Roediger, 2008; McDaniel, & Masson, 1985; Myers (1914); Storm, Bjork, & Storm, 2010; Wenger, Thompson, & Bartling, 1980; Whitten & Bjork, 1977). The act of retrieval is known to slow and otherwise interfere with forgetting of learned information (Hogan & Kintsch, 1971; Izawa, 1970; Maddox & Balota, 2015; Runquist, 1986; Wheeler & Roediger, 1992). Retrieval practice may further often constitute more transfer appropriate processing for many skills (Kolers & Roediger, 1984; McDaniel, Friedman, & Bourne, 1978; Morris, Bransford, & Franks, 1977), such as when the meaning of an L2 word must be retrieved during comprehension of the second language input. Because most use of acquired knowledge involves retrieval of various aspects of learned material as well as of their interrelationships, according to transfer appropriate processing theory (Morris et al., 1977), retrieval practice may promote such subsequent retrieval to a greater extent than practice that does not involve retrieval.

The retrieval effect is closely related to the testing effect, which is the widely observed finding that taking a test on the to-be-learned material is a more potent learning event than restudying the material, particularly for long term retention (Allen, Mahler, & Estes, 1969; Carpenter, Pashler, & Vul, 2006; Carrier & Pashler, 1992; Hogan & Kintsch, 1971; Kuo & Hirshman, 1996; Roediger & Butler, 2011; Roediger & Karpike, 2006; Spitzer, 1939; Thompson, Wenger, & Bartling,1978; Wheeler, Ewers, & Buonanno, 2003). The effects of testing have been obtained even in situations where there is no feedback following learners' attempts at retrieving information (Balota, Duchek, Sergent-Marshall, & Roediger, 2006; Hogan & Kintsch, 1971). Testing effects are still observed when processing time between a tested and a study-only condition is equated or is in favor of the restudy condition (Carpenter et al, 2006; Glover, 1989), indicating that the act of retrieving information is a cognitive process that differs fundamentally from simple study or exposure to the target material. Thus, the benefit of retrieval cannot be reduced to additional time on task (Carrier & Pashler, 1992; Kuo & Hirshman, 1996; Roediger & Karpike, 2006).

The terms *testing effect* and *retrieval effect* are often used interchangeably in research on their effects and underlying causes for their benefits. Further, it is widely believed today that the effects of testing are primarily due to retrieval processes that act on memory traces by elaborating and strengthening them (Bjork, 1975; Glover, 1989; Kornell, Hays, & Bjork 2009; McDaniel & Masson,1985; Roediger & Karpike, 2006). The effects of testing have been shown to increase with repeated testing (Karpicke & Roediger, 2008; Soderstrom et al., 2016; Wheeler & Roediger, 1992) and with feedback provided after retrieval attempts (Cull, 2000; Pashler, Cepeda,Wixted, & Rohrer, 2005). Further, unsuccessful retrieval attempts are still known as powerful learning events (Donaldson, 1971; Izawa, 1970; Kornell et al., 2009) and are known to promote deeper processing or encoding of the information contained in the feedback that follows than when the presentation of the same information is not preceded by a retrieval attempt. This is known as test-potentiated learning (Arnold & McDermott, 2013; Hays, Kornell, & Bjork, 2013; Izawa, 1970; Kornell et al., 2009; Roediger & Karpike, 2006).

Retrieval effort is argued to underlie the benefits of testing as well as findings that tests involving recall or constructed response lead to better subsequent retention than tests that only require easier tasks such as recognition or identification (Gardiner et al., 1973; Jacoby, 1978; Rowland, 2014). Retrieval effort is generally known to be beneficial for learning (Benjamin et al, 1998; Gardiner et al., 1973), and retrieval practice is known to be more beneficial the more effortful or complete the retrieval (Bjork, 1975; Glover, 1989; Whitten & Bjork, 1977). In fact, even when effort leads to more retrieval failures or errors during the learning phase, this still leads to better retention in the long term (Pashler, Zarow, & Triplett, 2003; Schmidt & Bjork, 1992; Soderstrom et al., 2016; Storm et al., 2010).

One way to induce more effortful retrieval is to put more time between the encoding event and the retrieval event (Cull, 2000; Glover, 1989; Jacoby, 1978; Modigliani, 1976; Roediger & Karpicke, 2006b; Soderstrom et al., 2016; Whitten & Bjork, 1977). Such a delay of retrieval has been shown to enhance learning from tests (Jacoby, 1978; Modigliani, 1976) and is attributed to greater effort required to retrieve information after some time has gone by since the encoding event. Extending the concept of fuller or more complete encoding that is argued to underly the benefits of spaced study relative to massed study, Glover (1989) argued that spaced retrieval attempts involve fuller retrieval of information than massed retrieval attempts because spaced retrieval is not supported by residual activation of the target stimulus as is the case when information is retrieved from short-term memory in massed retrieval. Such completeness of the retrieval process, in turn, leads to better memory for the studied material in spaced relative to massed retrieval. Thus, unlike retrieving information that was only recently presented and that still resides in short-term memory, retrieving information that was presented longer ago is more difficult, requires a more complete retrieval operation, and is, consequently, a more powerful learning method. Spaced retrieval practice has been widely found to be superior to massed retrieval practice (Craik, 1970; Cull, 2000; Cull et al., 1996; Logan & Balota, 2008).

Another way to ensure more difficult retrieval is to increase contextual interference (Bjork, 1994; Storm et al., 2010). This means that retrieval is more difficult when there is more similarity among the numerous learning targets or learning occurs amidst a multitude of other similar forms that a participant is exposed to even if these are not the focus of learning. Such high interference is usually characteristic of L2 learning contexts, where, the input contains large numbers of forms that often resemble each other and many of which follow the same phono- or orthotactic patterns. Particularly for novice learners, input can be overwhelming when it contains multiple unknown (and often not targeted in initial stages) forms, which may create interference. In a similar vein, interleaving retrieval attempts for different target items produces superior learning in the long term, which is attributed to more retrieval difficulty resulting from the interference of intervening retrieval attempts (Linderholm, Dobson, & Yarbrough, 2016).

Just as is the case with the spacing effect, retrieval practice produces benefits of considerable size and is a very general and consistent finding. Just as is the case with the spacing effect, its full potential has not been used in education (McDaniel & Fisher, 1991; Roediger & Karpicke, 2006a). Given that retrieval practice improves learning and that repeated retrieval attempts may further increase learning gains (Bahrick, 1979), a good question is how these retrieval attempts should be optimally distributed to achieve maximum learning. Spaced retrieval practice combines the benefits of spacing and retrieval and thus potentially maximizes learning. How best to do it is still a question (Storm et al., 2010), however.

In experiments that have directly measured study-phase retrieval success, study-phase performance has been shown to be consistently better in the massed condition than in the spaced condition, while the opposite holds for long term retention (e.g., Bahrick, 1979; Balota et al., 2006; Carpenter & DeLosh, 2005; Karpicke & Roediger, 2007). Similarly, in studies that have compared expanding schedules with uniform-interval schedules, acquisition performance is usually better in an expanding schedule (e.g., 1-3-5) than in an uniformly-spaced schedule (e.g., 3-3-3); however, performance on posttests that are administered with a longer delay is usually either equal in the two conditions or in favor of the uniformly-spaced condition (Balota et al., 2006; Carpenter & DeLosh, 2005; Logan & Balota, 2008; Storm et al., 2010). This seems counter-intuitive as the main rationale behind using expanding spacing schedules is that such a schedule supports successful study-phase retrieval at ever-increasing intervals, which is argued to underlie the beneficial effects of spacing. Further, in later repetitions that follow an expanding schedule, target items are retrieved after intervals that are considerably longer than those in the uniform-interval condition (because the average spacing is usually equated between the two conditions), which should further promote more effortful successful retrieval in an expanding schedule. Advantages of uniformly-spaced schedules over expanding schedules are often obtained in the absence of feedback, which means that information that is not retrieved during the study phase in the uniform-interval condition is simply forgotten. However, in terms of delayed posttest scores this condition still outperforms an expanding schedule condition that is specifically designed to minimize forgetting during the study phase. This finding is puzzling. It has been proposed that the initial retrieval attempt must be effortful to produce memory benefits (Karpicke & Roediger, 2007; Logan & Balota, 2008; Modigliani, 1976), which may explain why uniformly-spaced schedules (where the initial retrieval attempt is always after a longer interval than is the case in an expanding schedule) do no worse and often even better than expanding schedules, where retrieval success is higher during study. In fact, the benefits of equal-interval schedules have been attributed by some researchers to less retrieval success during acquisition under such conditions (Storm et al., 2010). This suggests that study-phase retrieval success may play a limited role under certain circumstances, particularly when such retrieval is less effortful (Pashler et al., 2003; Storm et al., 2010).

Retrieval practice is usually investigated within a paired-associate learning format. PAL consists of learning to associate two members of a pair of stimuli (Allen et al., 1969; Carrier & Pashler, 1992; Cull et al., 1996; Greeno, 1964; McDaniel & Masson, 1985; Nelson, Leonesio, Shimamura, Landwehr, & Narens, 1982). The task of retrieving the second member of a pair of associates is also sometimes referred to as a cued-recall task (e.g., Carpenter, et al., 2006; McDaniel & Masson, 1985). This task is relevant for many learning situations, such as for learning to associate a meaning with a foreign word, and is a method that is often used in L2 vocabulary learning. In the field of psychology, the studied pairs are most often two weakly related L1 words (e.g., Jacoby, 1978; Logan & Balota, 2008; Maddox & Balota, 2015). While useful for the investigation of many memory phenomena, the task of associating two L1 words is not in itself a real-life task. More real-world learning targets have also been used, such as the learning of low-frequency L1 words with their definitions (Gardiner et al., 1973; Rohrer, Taylor, Pashler, Wixted, & Cepeda, 2005, Exp. 2), or L1-L2 or L2-L1 translation pairs (Arnold & McDermott, 2013; Barrick, 1979; Bahrick, Bahrick, Bahrick, & Bahrick,1993; Callan & Schweighofer, 2010; Carrier & Pashler, 1992; Kang, Lindsey, Mozer, & Pashler, 2014; Karpicke & Roediger, 2008; Pashler, et al., 2005; Pashler et al., 2003; Pavlik & Anderson, 2005; Pyc & Rawson, 2009). Psychology studies using foreign word learning have generally obtained benefits of spaced retrieval practice over massed retrieval practice as well as benefits of retrieval over restudying, particularly on delayed tests, which reflect long-term knowledge that is more relevant for L2 learning.

Research into the spacing effect and retrieval practice in second language acquisition

The spacing effect has generated some interest in the field of second language acquisition. A small number of studies have looked at the effects of spacing practice on L2 grammar acquisition (Bird, 2010; Miles, 2014; Kasprowicz, Marsden, & Sephton, 2019; Rogers, 2015; Suzuki, 2017; Suzuki & DeKeyser, 2017; Suzuki, & Sunada, 2019). Other studies have explored the effect in the context of vocabulary acquisition (Bahrick et al.,1993; Bahrick & Phelps, 1987; Bloom & Shuell, 1981; Miles & Kwon, 2008; Nakata, 2015; Nakata & Suzuki, 2018; Nakata & Webb, 2016; Schuetze, 2015). Thus, for instance, Nakata (2015) investigated the effects of spacing study of vocabulary within a PAL format. More specifically, he investigated the effects of an expanding spacing schedule. Recall that an expanding spacing schedule refers to using increasingly longer time intervals between repetitions (Kang et al., 2014; Landauer & Bjork (1978) rather than constant intervals. Nakata found a large positive main effect for spacing but only a small positive effect of using an expanding schedule.

Similar results were obtained in Schuetze (2015), where in a between-subjects design, students studied English-German translation pairs in a classroom setting. The translation pairs were presented four times in total for 8 seconds per presentation with a different number of days between the repeated presentations. Participants were tested for production of the L2 German words cued by their L1 English translations three times, with the last test being eight weeks after the study phase. Schuetze found that results from the expanding-interval schedule practice were superior to those for the equal-interval schedule in the shorter term while this pattern was reversed in the longer term, where the equal-interval group showed much less forgetting than the expanding-interval group. This is in line with findings in psychology. An important difference between immediate and four-day delayed posttests was also found by Bloom and Shuell (1981) in another between-subject classroom study, where L1 English learners studied L2 French words in written vocabulary activities. The words were practiced either within one session (the massed condition), or distributed over three days (the spaced condition). While similar levels of learning gains were obtained in the massed and spaced conditions on the immediate posttests, on the delayed test administered seven days later, the scores in the spaced study condition were superior to those in the massed study condition.

Some SLA research into spacing and vocabulary learning has also included investigations of other variables that are relevant for vocabulary learning contexts. Thus, Nakata and Suzuki (2019) investigated the effects of spaced practice on the acquisition of semantically related and unrelated words, also in a PAL format. Because learning of semantically related words (semantic clustering) had been found in previous research to produce interference effects that hinder acquisition, the authors reasoned that spacing practice of semantically related words would alleviate such interference and would, therefore, be beneficial for learning of semantically related words. Thus, the authors asked whether spacing practice benefits semantically related and unrelated words differently. The authors found that spacing was beneficial for both related and unrelated words and that, contrary to expectation, unrelated words benefited from spacing more than did related words. Nakata and Webb (2016) manipulated learning set size, or the number of words studied at one time, (Experiment 1) and spacing (Experiment 2) in learning of low-frequency L2 English words in a PAL format with retrieval practice, where participants were to produce the second member of a pair (both L2-L1 and L1-L2 translation for Experiment 1 and only L1-L2 translation for Experiment 2) before being provided with feedback. The authors found that spacing had larger beneficial effects than did the size of the learning set.

Thus, for the most part, second language vocabulary learning studies have shown that spacing repeated study is beneficial. However, some second language studies have reported no effect of spacing or even the opposite effects, where spacing was found to be actually detrimental to learning outcomes. Thus, Elgort and Warren (2014), who investigated novel word learning from incidental exposure during reading of a long authentic text (without the use of a dictionary) over a ten-day period, found that novel words that repeated in the same chapter of the book were remembered better than those that repeated across chapters, especially for the less proficient readers. The authors speculate that this may be due to memory trace decay between repetitions, which may interfere with the development of lexical semantic representations and abstraction of a core meaning of a word. The fact that the more widely spaced repetitions were particularly detrimental for the lower proficiency learners is in line with the argument that memory trace survival is important. Retrieving the previous encounter with a word or processing an encounter with a given word as a repetition may be less likely to be successful if the process of L2 comprehension is a difficult task (Bui et al., 2013).

Similarly, Suzuki and DeKeyser (2017) found no advantage of practice separated by a week over practice repeated by a day for proceduralization of grammatical knowledge (and, in fact, found some benefit for the latter). The authors attribute this finding to the fact that the task used in their study was more complex compared to psychology experiments that have used simple tasks and showed large benefits of spacing. Indeed, optimal ISI is known to be shorter for more complex tasks (Donovan & Radosevich, 1999), which, again, makes sense if one assumes that the memory traces that are established need to be strong enough to survive longer lags and that any interference produced by a complex task that is performed in the interim may decrease the chances of retrieving prior encounters at a subsequent repetition (Bui et al., 2013; Verkoeijen et al., 2005).

Detrimental effects of distributing second language study have also been obtained in the context of program evaluation. Such research has compared the effectiveness of intensive programs, where study sessions are massed closely together, with extensive programs, where study sessions are spread more widely over time. This research has consistently shown that intensive programs are more effective (Collins et al., 1999; Serrano, 2011; Serrano & Munoz, 2007; White & Turner, 2005), particularly for lower-proficiency learners (Serrano, 2011). This, again, is contrary to the widely observed benefits of distributing practice documented in the field of psychology and constitutes a finding of a reverse effect. Such a reverse finding suggests that effects of distributing practice may depend on variables that need to be taken into account and whose effects need to be known (Rogers, 2017). Some have attributed the failure to obtain a spacing effect in this research context to the simple fact that these studies did not use a delayed posttest (Bird, 2010; Serrano & Munoz, 2007), where the spacing effect usually manifests itself much more strongly (Rawson & Kintsch, 2005). Others have stressed that the type of knowledge targeted and the context of acquisition of this knowledge may be different or more complex in a language learning context than what is widely used in psychology experiments. It is argued, therefore, that applying findings from psychology studies to language learning contexts is not always straightforward (Bird, 2010, p. 640; Rogers, 2017). If we assume that processing repetitions as repetitions is important for learning from spaced practice, it may also be the case that detrimental effects of spacing on learning in extensive programs comes from the fact that it is more difficult to retrieve material presented in a previous session, or each new session may not have a high reminding potential of the previous session, when it is separated from the previous session by a longer time interval. While there may be some overlap between consecutive sessions, this may be more clearly felt when the sessions occur closely together than when they are separated by longer periods of time, allowing many of the details that might be used as cues for retrieval of previous encounters to fade to a greater extent.

SLA research has, thus far, focused mainly on the question of whether or not distributing practice produces superior learning outcomes for different aspects of a second language, without much direct investigation of the underlying mechanisms. The expectation that spaced practice should be beneficial for learning is based on the ubiquitous finding of benefits of spacing in psychological research. However, as discussed above, applying findings from psychology to L2 learning and teaching situations may not always be straightforward (Rogers, 2017). Further, it is widely believed today that beneficial effects of spacing study may rely on an interplay of different underlying mechanisms depending on the learning situation or target task (Gerbier &Toppino, 2015; Glenberg & Smith, 1981; Greene, 1989; Kornell & Bjork, 2008; Russo & Mammarella, 2002). The operation of these different mechanisms may further be affected by variables that characterize specific learning contexts (Verkoeijen et al., 2004; Verkoeijen et al., 2005). It is, therefore, important to investigate the process as well as the product of second language study under different levels of spacing. Only a few SLA studies have attempted an investigation of the process itself, however. Nakata and Suzuki (2019), for instance, measured learners' retrieval success during the study phase through the task of overt L2-L1 translation. This methodology is similar to the one used by Maddox and Balota (2015), who asked their participants to retrieve the second member of a paired associate. In addition to using learning targets that are more relevant for SLA, an important difference that also makes Nakata and Suzuki's study more relevant for L2 learning is that they provided feedback to the learners after each retrieval attempt. However, Nakata and Suzuki did not investigate posttest performance as a function of successful study-phase retrieval and, therefore, cannot inform as to the potential mediating effects of study-phase retrieval success. Further, in order to avoid a large number of unsuccessful retrieval attempts by their participants, they broke down study of their 48 target words into two sets of 24, thereby avoiding a situation where the effects of study-phase retrieval failure on learning outcomes could be directly tested. Suzuki and DeKeyser (2017) included an ad hoc analysis of lexical retrieval performance during training on an element of L2 Japanese morphology. The distributed practice group, who practiced in two sessions separated by a week (versus one day, which was the case for the massed group), had more difficulty retrieving the vocabulary during the second session. The authors considered this variable ad hoc and speculated that ease and success of lexical retrieval may affect the nature of cognitive processes involved in distributed and massed learning.

Another study that investigated the process as well as the product of learning under differential spacing is Koval (2019), in which I used eye-tracking methodology to test the deficient processing account of the spacing effect in L2 vocabulary learning from sentence reading. Two levels of ISI were used: the target words appeared either in consecutive sentences or in sentences that were separated by other sentences containing other target words plus a six-minute distractor math task. The choice of account was motivated by proposals in the field of SLA that attentional processing benefits learning of a second language in general (Gass, 1988; Robinson, 2003; Schmidt, 1990) and vocabulary learning success in particular (Godfroid et al., 2018; Godfroid, et al., 2013). I found that reading times on the target words decreased with repeated encounters for both spaced and massed repetitions (as had been found in other studies of L2 vocabulary learning from reading, Godfroid, et al., 2013) but did so more dramatically in the massed condition, resulting in less overt visual attention given to repeated encounters with the target vocabulary that occurred in consecutive sentences. I further found that attentional processing, as measured by total reading time, was a significant mediator for the beneficial effects of spacing that were observed in the study, confirming that an attentional account of the spacing effect has relevance for contextual second language vocabulary learning. In this study, target words were embedded in different sentence contexts. Different contexts have previously been shown to benefit massed repetitions but to have a detrimental effect on learning from spaced repetitions (Verkoeijen et al., 2004). This finding has been explained in terms of a higher chance of failure to recognize a word as repeated (failure of study-phase retrieval) when it repeats in different contexts and the repetitions are widely spaced. To investigate whether differences in the sentence contexts may have detracted from learning in the spaced condition, I additionally investigated the downward trajectory in reading times in the spaced repetitions for evidence of a repetition effect (Joseph, Wonnacott, Forbes, & Nation, 2014; Pellicer-Sánchez, 2016; Rayner & Duffy, 1986; Rayner, Raney, & Pollatsek, 1995). I used first exposures in the massed condition that occurred across the four blocks of the study phase as controls for potential effects of order or fatigue, thus isolating the effects of repetition from order effects. I found that there was significant facilitation in the total reading time measure that came with repeated encounters, suggesting that repeated encounters in the spaced condition were, in fact, mostly processed as repetitions despite differences in sentence contexts. Such an investigation of a repetition effect in terms of facilitation in reading times constitutes an indirect memory test (Richardson-Klavehn & Bjork, 1988), one in which participants are not asked to provide an overt retrieval response – as was the case in the explicit repetition detection judgments in studies such as Bellezza et al., (1975) and Maddox et al., (2018) or retrieval of the paired associate in Maddox & Balota (2015). In my study, intentionality of learning (Verkoeijen et al., 2005) combined with the relative ease of the intervening task (L1 sentence reading and simple math operations), which means relatively low levels of interference (Bui et al., 2013), may have aided successful study-phase retrieval across the spaced encounters.

More research is needed that explores the process as well as the product of learning second language material under different levels of temporal distribution of repetitions. SLA research needs to explore the potentially relevant mechanisms that may underly any effects of spacing as well as how the operation of the mechanisms may be affected by variables that are relevant for SLA contexts. The present study sets out to test the predictions of the dual-mechanism reminding account (Benjamin & Tullis, 2010) by exploring the contribution of study-phase retrieval success and effort to learning L2 vocabulary from repeated exposures at three different levels of within-session ISI in a PAL format. The focus on a dual-process account that includes successful study-phase retrieval as an underlying mechanism for this investigation is motivated by the fact that current theories of the spacing effect include study-phase retrieval as an important element in learning from repetition and a crucial precondition for observing beneficial effects of spacing. It is further motivated by the fact that a failure to process repeated encounters with target items as repetitions has been cited in SLA research as a potential explanation for failures to observe benefits of spacing (see, e.g., Elgort & Warren, 2014; Serrano, 2011). The inclusion of the second element of effortful processing is motivated by the widely-held belief that attentional engagement and effort are beneficial for learning of second language vocabulary (Godfroid et al., 2013; Laufer & Hulstijn, 2001; Mohamed, 2018; Schmitt, 2008) as well as my finding that deficient processing of massed encounters mediates the benefits of spacing in L2 vocabulary learning (Koval, 2019).

Both success and effort of retrieval at repetition may depend on a number of factors. One such factor is likely to be the length of time a learner spends studying a word per repetition. The more time a learner spends studying such a word the stronger the resulting encoding is likely to be, which may be more likely to survive longer ISIs (Verkoeijen & Bouwmeester, 2008) and thus promote retrieval success at a subsequent repetition. Further, the longer a word is studied with its meaning, the less effort may be required for retrieval of the meaning at a subsequent repetition. Thus, study time may have important effects on the operation of both underlying mechanisms tested in the present study.

Research questions

The aim of the present study is to test the contribution of the dual mechanism of effortful successful retrieval (Benjamin & Tullis, 2010) to any effects of lag on learning second language vocabulary from L2-L1 retrieval practice in a PAL format. Another aim is to test any effects of study time on the operation of the two proposed mechanisms as well as on learning outcomes. The present study is motivated by the following research questions:

1. Does the amount of lag between repeated retrieval attempts affect learning from retrieval practice in a PAL format, as measured by immediate and delayed form-recognition and translation posttests? Does the amount of time given for study of an L2-L1 translation as feedback affect this relationship?
2. Does the amount of lag between repeated retrieval attempts affect study-phase retrieval effort and success? Does the amount of time given for study of an L2-L1 translation as feedback affect this relationship?
3. Does the dual mechanism of successful effortful retrieval mediate effects of spacing? Is the operation of the two mechanisms affected by the amount of time a learner is given to study an L2-L1 translation pair per repetition and in total?

CHAPTER 3

METHOD

Participants

Fifty-two native speakers of American English (healthy young adults) participated in the experiment. These were mostly undergraduate students in a wide variety of majors at Michigan State University who had responded to an ad about the study that had been placed through the Office of the Registrar. Twenty-two were male and 28 were female, ages 18-29 (*M* = 20.04, *SD* = 2.08, Median = 20). Most of these students had studied at least one foreign language, with the number of foreign languages varying from one to four (*M* = 1.72, *SD* = 0.81, Median = 2). Proficiencies ranged from 1 to 5 (*M* = 2.31, *SD* = 1.03, Median = 2) on the self-assessment question that ranged from 1 (lowest) to 5 (highest). Spanish French, and German were the most frequently indicated as languages studied by the participants. Other languages studied included Chinese, Japanese, Korean, Russian, Thai, ASL, Burmese, Italian, Arabic, Polish, and Greek. However, none of the participants were familiar with the Finnish language. One participant reported having travelled to Finland; however, he was not familiar with the language beyond one name of a dish, as he said. Another participant reported that his grandfather is from Finland. However, the participant reported to have no knowledge of the language. Participants' responses to the first encounters with the target words will be further used in this study as a kind of pretest to ensure no prior knowledge of the target words. A question on the background questionnaire administered after the study phase will further explore participants pre-existing familiarity with the target words.

Materials and design

I used a fully counterbalanced within-item within-participant design. The experiment consisted of a study phase, a distractor math task, 30-minitue delayed vocabulary posttests (referred to as immediate posttests) that measured form recognition of the target words as well as participants' ability to produce and select their L1 English translations, one- to two-weeks delayed vocabulary posttests (referred to as delayed posttests) that were identical to the immediate posttests except for item order randomization within and between participants, and a linguistic background questionnaire.

Study phase. I selected Finnish as the target language for the study. The use of an existing language, where each word is paired with its actual English translation, was deemed to be more ecologically valid. Finnish is a relatively uncommon language in the US, which minimizes the chance of prior exposure among American students (the target participant population). Being a language of the Finnic family, it also bears little resemblance to English or languages that are commonly studied by US students. Further, Finnish is written in the same alphabet as English, the participants' L1, which allows to control for reading difficulty.

Seventy-two simple generic Finnish nouns with all diacritic marks removed were chosen as the target words for this study. None of these nouns were cognates of their English translations. The 72 words were divided into two main lists (36 words each). The words on each list served as experimental repeated targets half of the time, and as once-presented controls the other half. The purpose of the unrepeated controls was to investigate the effects of retrieval practice in the three ISI conditions against a baseline of no practice beyond one study event. Within each of the two lists, the words were further divided into three ISI lists (12 words each), each to be used in each of the three levels of ISI (massed, short-spaced, and long-spaced) when serving as experimental items. A rotation was performed on the items for counterbalancing. Each time the repeated items were changed from one ISI condition to the next, the control items changed place in terms of their order within the experimental sequence. This way, each control item got to appear at the beginning, in the middle, and toward the end of the experimental sequence. Each ISI list (12 words) was further divided in half for the two levels of study time (3 vs. 9 sec). This was done such that words in the two study time lists were matched on the number of letters. Thus, each of the two levels of study time was equated on the number of words and the number of letters per word; it also had each condition equally represented. I further counterbalanced the words in terms of study time. Four-five participants fell into each of the 12 resulting counterbalancing lists.

The target words ranged in length from four to eight letters. On both lists, each ISI sublist contained two four-letter words, four five-letter words, three six-letter words, one seven-letter word, and two eight-letter words (see Appendix A for a list of the target Finnish words with their English translations). The N-Watch program (Davis, 2005) was used for information on frequency of the English translations. CELEX frequency and LOG 10 frequency were used. In N-Watch, LOG 10 frequency is based on the CELEX English Linguistic Database (Baayen, Piepenbrock, & van Rijn, 1995). The reason for including LOG 10 transformed indices is the fact that the relationship between word frequency and psycholinguistic measures such as lexical decision time is known to follow a logarithmic function (Davis, 2005). This refers to the fact that the frequency difference between any two low-frequency words has been found to have a larger effect on psycholinguistic measures such as reaction time than the same difference between two high-frequency words. Brysbaert, Warriner, and Kuperman’s (2014) database of concreteness ratings was used for indices of concreteness. The English translations ranged from 0.43 to 2.63 on their LOG10 frequency and from 3.3 to 5 on their concreteness values. The target nouns were matched exactly on the number of letters between the two lists and also among the three ISI sublists within each such list. The resulting lists were further roughly matched on indices such as frequency and concreteness (see Appendix B for frequency and concreteness information for the English translations in the two main lists as well the three sublists within each list). Two hundred and ten additional Finnish words were selected to serve as practice and recency items as well as filler trials during the study phase. Some of these repeated and others were only presented once. Some of these were followed by their translations and others were not. The filler items were similar to the target items in terms of structure (the same overall length and orthotactic patterns, as would be expected among words from the same language).

A practice block preceded the experimental sequence. A recency block followed the sixth experimental block. These blocks contained many of the same fillers that were used in the study phase. None of the target words were used in the practice block or the recency block. The purpose for the recency block was to minimize any recency or order effects on the 30-min delayed (immediate) posttest for words that occurred later rather than earlier in the experimental sequence. Fillers that were associated with their L1 translations were not in any way different from the target words from the point of view of the participant. Further, these often repeated in a similar pattern to the target words, except that the number of repetitions and the pattern of repetition was different and more haphazard. This was done to prevent participants from anticipating a pattern of repetition for the target items. The practice block served to minimize any effects of primacy on the target items that were introduced at the beginning of the study phase as well as to familiarize the participant with the procedure.

During the experimental portion of the study phase, the target words were studied in six experimental blocks. The words in the massed condition repeated six times within each block. These were separated by 0-1 intervening trials (1 second in the case of zero intervening trials: here the interval refers to the time between the offset of the Finnish word presented with its translation and the onset of the next corresponding trial, where only the Finnish word is presented on the screen until a response is made; or, in the case of one intervening trial, 5-21 seconds, depending on the speed of response to the filler item). The intervening Finnish words that separated massed repetitions were always fillers and were never accompanied by a translation in order to preserve the massed nature of study. The words in the short-spaced condition repeated six times over two consecutive blocks (three times per block) and were separated by 17-38 trials within a block and by 12-22 trials plus the 6 minute-distractor math task between two adjacent blocks (3-4 or 6-8 minutes between repetitions). The words in the long-spaced condition repeated once per block and were separated by 71-119 trials plus the six-minute intervening distractor math task (16 -19 minutes between repetitions). The average position across the experimental sequence was equated for the words in all four conditions (massed: 249.82; short-spaced: 249.97; long-spaced: 251.10; controls: 248.44) and was not different statistically, $F(3, 13104) = .159$, $p = .924$. *Figure 1* presents graphically the conceptual pattern of repetition for one item in each of the three ISI conditions across the six blocks.

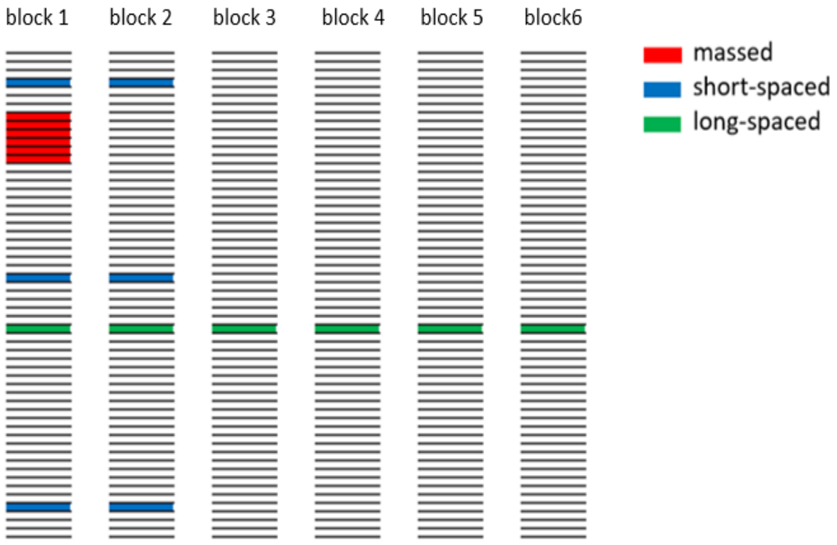
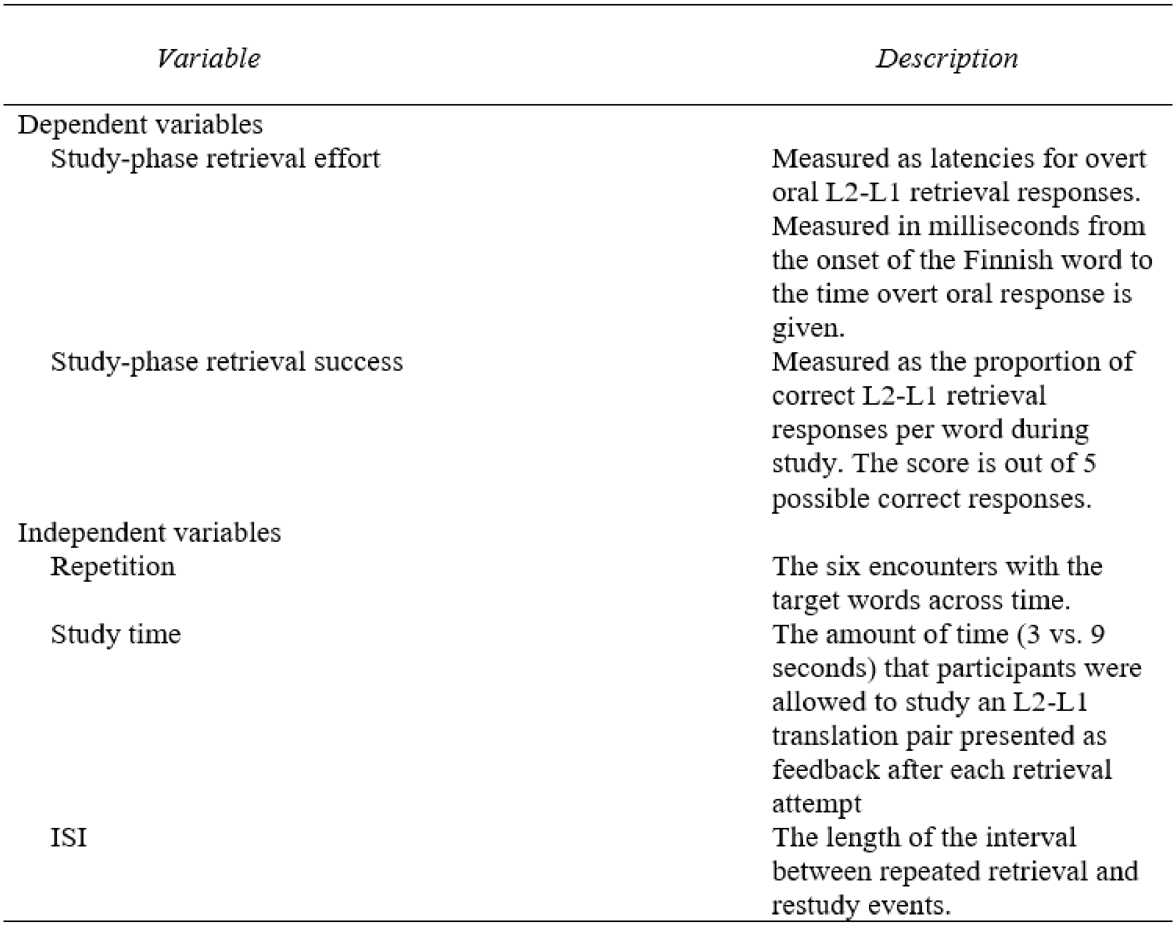


Figure 1: A conceptual illustration of the repetition pattern for one item

Each experimental block started and ended with three filler items. Further, the conditions were equally represented at the beginnings and ends of blocks: blocks 2, 4, and 6 began and ended with two control items; block 1 began and ended with a massed item (all six repetitions); block 3 began and ended with two short-spaced items (1 repetition); block 5 began and ended with two long-spaced items (1 repetition). The reason for one item in the massed condition beginning and ending a block was because six repetitions had to be consecutive in this condition. It was hoped that using one repetition of two different items in the other conditions would offset this difference. Table 2 presents the variables used in the latency analysis.

Distractor math task. A simple math task was performed for six minutes between the six blocks as well as between the final sixth block and the recency block. During this time, participants were given multiplication, addition, subtraction, and division tasks to perform. Participants did both mental math and math that they wrote out on paper to ensure variety in the activity and minimize boredom and fatigue.

Table 2: *Variables used in the study-phase analyses*



Variable	*Description*
Dependent variables	
Study-phase retrieval effort	Measured as latencies for overt oral L2-L1 retrieval responses. Measured in milliseconds from the onset of the Finnish word to the time overt oral response is given.
Study-phase retrieval success	Measured as the proportion of correct L2-L1 retrieval responses per word during study. The score is out of 5 possible correct responses.
Independent variables	
Repetition	The six encounters with the target words across time.
Study time	The amount of time (3 vs. 9 seconds) that participants were allowed to study an L2-L1 translation pair presented as feedback after each retrieval attempt
ISI	The length of the interval between repeated retrieval and restudy events.

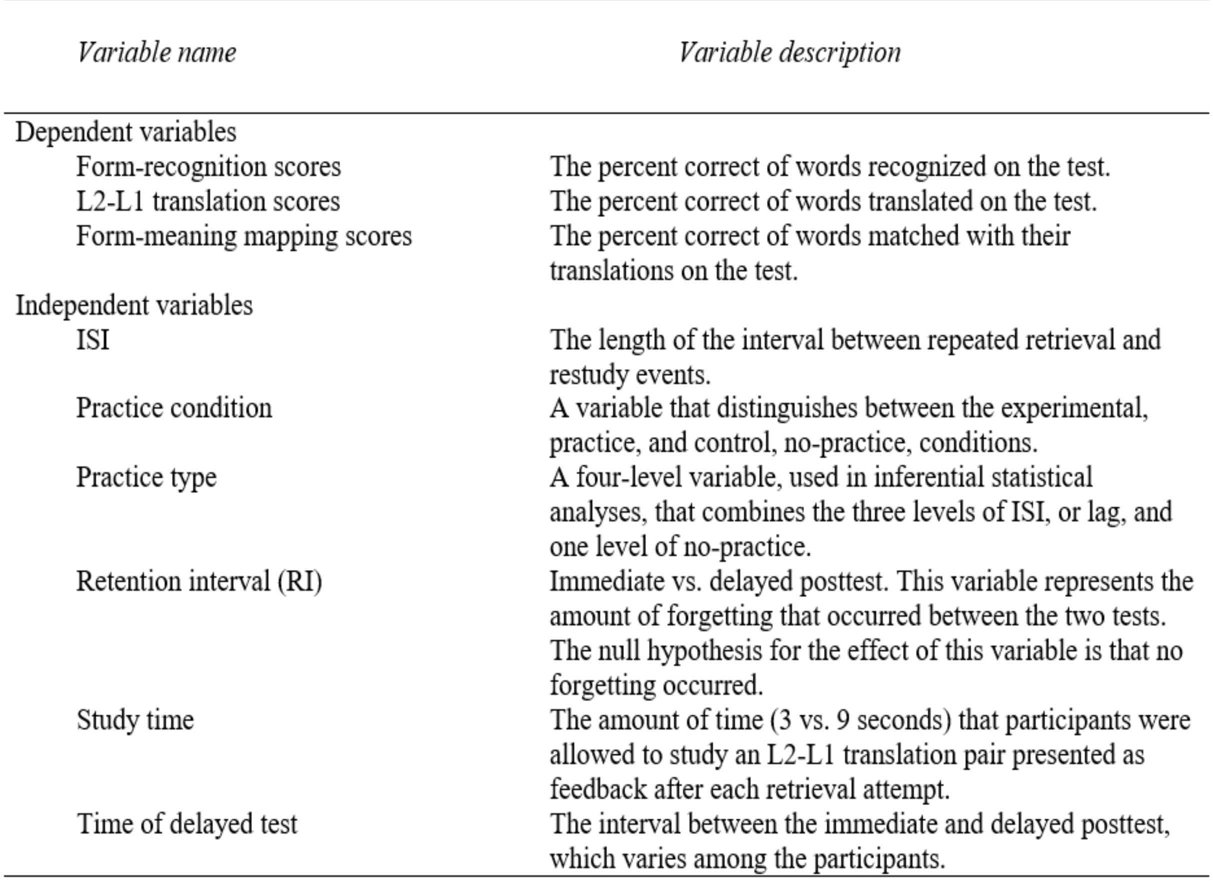
Posttests. Three identical (except for item order randomization) sets of immediate and delayed paper and pencil posttests were used to measure learning gains. In each of the two administrations, Posttest 1 was a form-recognition test. Here, the 72 target words were presented among 156 new Finnish words (distractors) that had not occurred during the study phase. Participants were to underline words that they recognized as ones studied during the study phase (see Appendix C for the instructions for this test and Appendix D for the test sheet). An effort was made to ensure that the distractors that appeared on the posttests were not too similar in form to the target words and to the distractors that were encountered during the study phase, particularly for distractors that appeared on the immediate posttest, as this posttest followed only 30 minutes after the study phase.

In each of the two administrations, Posttest 2 was an L2-L1 translation test. Here, participants were to write the English translations next to the target Finnish words (on Sheet A) presented without distractors (see Appendix C for the instructions for this test and Appendix E for the test sheet).

In each of the two administrations, Posttest 3 was a form-meaning matching test. Here, participants were presented with the English translations for all the target Finnish words (Sheet B). Participants were to add the number associated with each English translation on Sheet B next to the corresponding Finnish word on Sheet A, which had been used in Posttest 2 (see Appendix C for the instruction for this test and Appendix F for the test sheet). The Finnish word sheet from Posttest 2 was used here instead of a new sheet because participants had at this point familiarized themselves with the layout of the Finnish words, resulting in more ease of location of the words. Presenting them with a new sheet of Finnish words would have added unnecessary search for the words.

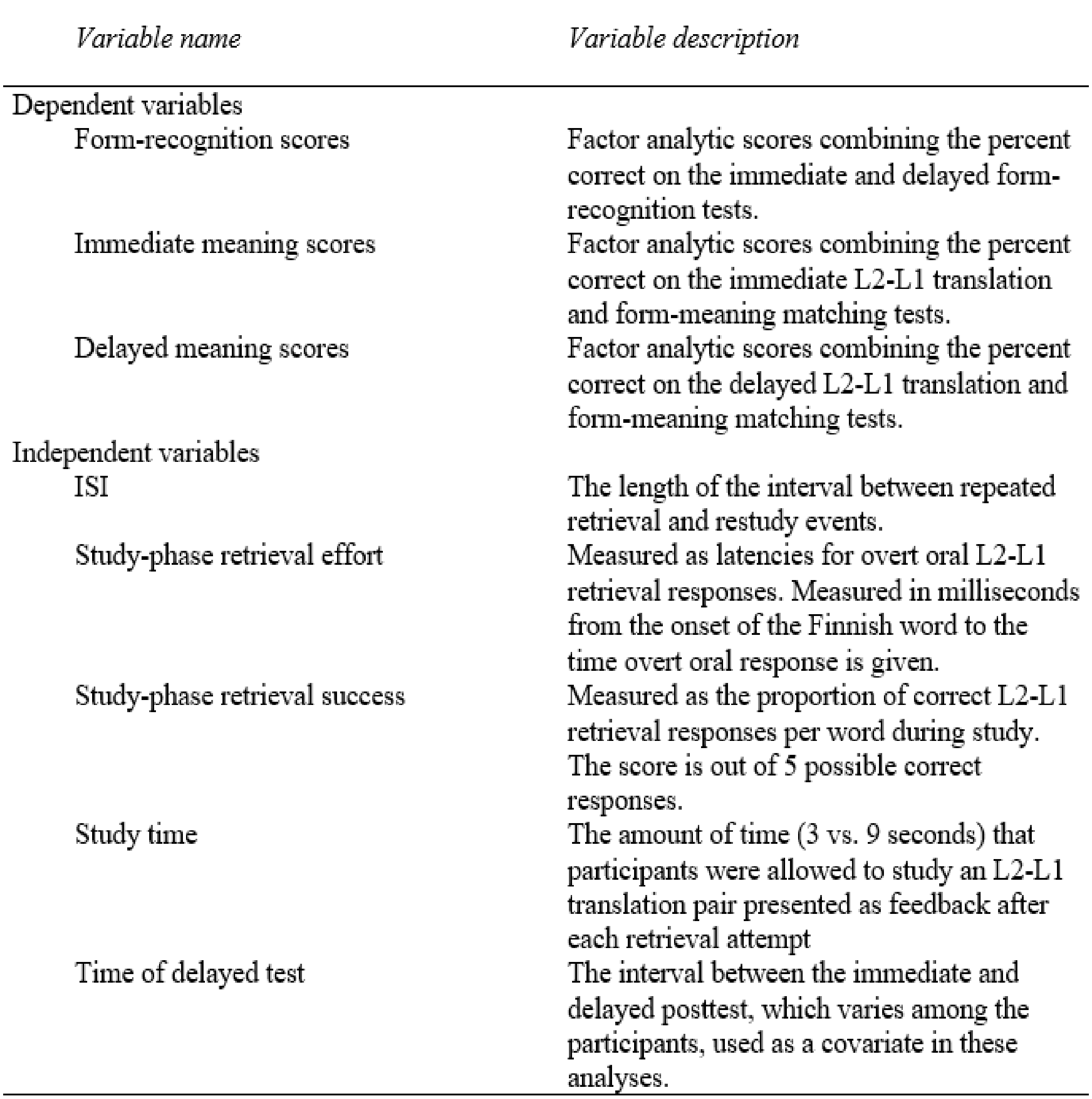
A different set of distractors was used for the immediate and delayed form-recognition tests. This was done to prevent participants from selecting an item on the delayed posttest due to the fact that they had seen it on the immediate posttest. All posttests were randomized in terms of order for each participant and also between the immediate and delayed administrations within each participant. Table 3 presents a summary of the variables used in the posttest analyses. Table 4 presents the variables for the mediation analyses.

Table 3: *Variables used in the posttest analyses*



Variable name	*Variable description*
Dependent variables	
Form-recognition scores	The percent correct of words recognized on the test.
L2-L1 translation scores	The percent correct of words translated on the test.
Form-meaning mapping scores	The percent correct of words matched with their translations on the test.
Independent variables	
ISI	The length of the interval between repeated retrieval and restudy events.
Practice condition	A variable that distinguishes between the experimental, practice, and control, no-practice, conditions.
Practice type	A four-level variable, used in inferential statistical analyses, that combines the three levels of ISI, or lag, and one level of no-practice.
Retention interval (RI)	Immediate vs. delayed posttest. This variable represents the amount of forgetting that occurred between the two tests. The null hypothesis for the effect of this variable is that no forgetting occurred.
Study time	The amount of time (3 vs. 9 seconds) that participants were allowed to study an L2-L1 translation pair presented as feedback after each retrieval attempt.
Time of delayed test	The interval between the immediate and delayed posttest, which varies among the participants.

Table 4: *Variables used in the moderated mediation analyses*



Variable name	*Variable description*
Dependent variables	
Form-recognition scores	Factor analytic scores combining the percent correct on the immediate and delayed form-recognition tests.
Immediate meaning scores	Factor analytic scores combining the percent correct on the immediate L2-L1 translation and form-meaning matching tests.
Delayed meaning scores	Factor analytic scores combining the percent correct on the delayed L2-L1 translation and form-meaning matching tests.
Independent variables	
ISI	The length of the interval between repeated retrieval and restudy events.
Study-phase retrieval effort	Measured as latencies for overt oral L2-L1 retrieval responses. Measured in milliseconds from the onset of the Finnish word to the time overt oral response is given.
Study-phase retrieval success	Measured as the proportion of correct L2-L1 retrieval responses per word during study. The score is out of 5 possible correct responses.
Study time	The amount of time (3 vs. 9 seconds) that participants were allowed to study an L2-L1 translation pair presented as feedback after each retrieval attempt
Time of delayed test	The interval between the immediate and delayed posttest, which varies among the participants, used as a covariate in these analyses.

Linguistic background questionnaire. A background questionnaire (see Appendix G) was used to collect information on participants' age, sex, any foreign languages studied, and any other information that the participant felt was relevant. The questionnaire also asked the participants to indicate whether any of the studied words had struck them as familiar upon initial encounter and to elaborate if the answer was yes.

Instruments

The DMDX software (Forster & Forster, 2003) was used on an *HP* lap top computer for stimulus presentation and recording of the response latencies. Two *Transcend* voice recorders were used to record participants' oral responses. All posttests and the background questionnaire were on paper. Microsoft Office 365 Excel was used for building and rotating the study-phase scripts as well as for randomizing posttest item presentation order and coding of the auditory responses.

Procedure

The experimental procedure is summarized in *Figure 2*.

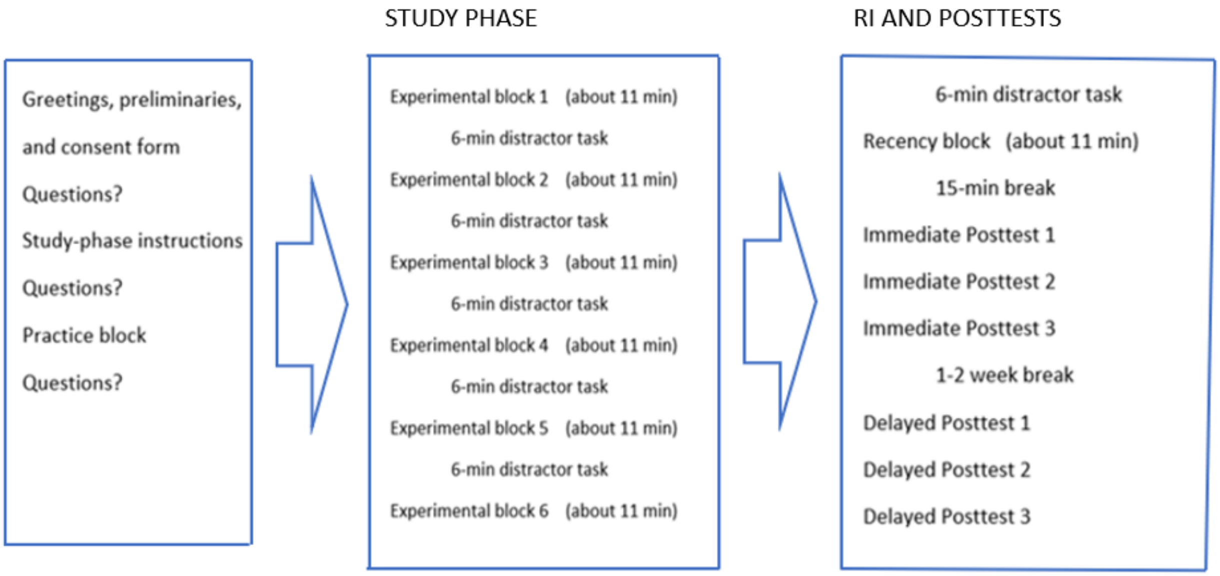


Figure 2: A summary of the experimental procedure

The entire experiment was approximately 3 hours 45 minutes in duration, over two sessions, per participant. Session one was about 3 hours and 10 minutes in duration. Session two was between 20 and 35 minutes in duration. Session one included the study phase, a 15-minute break, the immediate posttests, and the background questionnaire. Session two included only the delayed posttests. The two sessions were separated, depending on participant availability, by approximately one or two weeks. The experiment was conducted with each participant individually, in a small quiet lab. The researcher met with each participant at a time scheduled via email.

The experimental sequence was as follows. First, the participant read and signed the consent form. They also asked any questions that they had during the reading of the consent form. This was followed by reading of the instructions for the study phase from the computer screen (Appendix H). After and during reading of the instructions, the participants were encouraged to ask any clarification questions. This was followed by the practice block, which consisted of 83 trials. After and during the practice block, the participants were encouraged to ask any further questions they may have. Following the completion of the practice block, the experimental blocks were completed in order, separated by 6-minute distractor tasks. Block one consisted of 110 trials. Each subsequent block consisted of 90 trials. Block one took 12 minutes, on average, and each subsequent block took 11-12 minutes, on average, to complete.

Figure 3 presents an example of an experimental study-phase trial sequence.

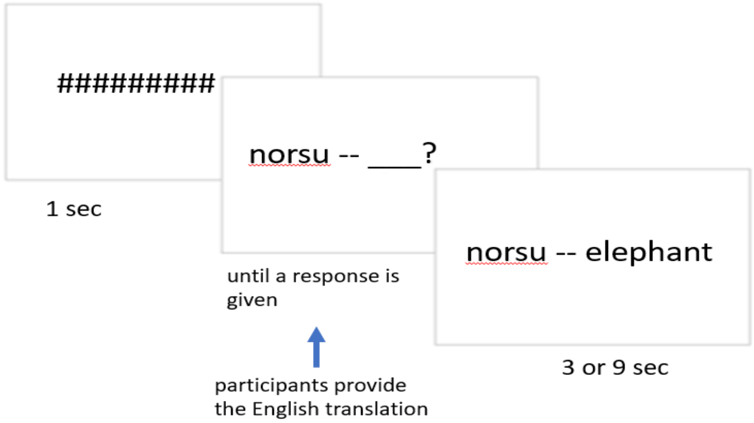

#########
1 sec
norsu -- ____?
until a response is given
participants provide the English translation
norsu -- elephant
3 or 9 sec

Figure 3: An example of one experimental trial sequence

Each trial started with the presentation of a row of hash marks (########) that stayed on the screen for 1 second and was replaced by a target Finnish word with a dash and an underscore with a question mark (norsu --- _________?) prompting the participants to produce the English translation for the word. The participants were to say these translations aloud while their responses were audio-recorded. If the participant could not remember the translation or if they believed that they had never seen the translation for a given word, they were to say “I don’t know”. Response time was recorded through a button press by the researcher (as in Maddox & Balota, 2015), which initiated the next screen, on which the Finnish word was presented with its paired associate L1 translation (norsu --- elephant). The pair stayed on the screen for either 3 seconds or 9 seconds, depending on the level of exposure duration assigned to the word for the specific rotation version, after which the next trial began. Distractor words that were presented with translations followed the same sequence. If a distractor word was not presented with a translation, the button press initiated the next trial. However, the next trial did not begin until the distractor had been on the screen for 3 seconds, which was held constant across all distractors that were not followed by a translation. A line of hash marks (#########) preceded the presentation of each word. This was used to signal the beginning of a new trial and a new word that was about to be presented.

At the end of each experimental block, the participants were asked whether they needed to step out. Whenever a participant indicated that they did, such as to use the bathroom or get a drink of water, they were allowed to do so before beginning the distractor math task. With these participants, the math task was cut a bit short, however, the break was a bit longer than 6 minutes to strike a balance between the loss in terms of the time spent on the cognitive activities involved in the math task and the gain in absolute time between the experimental blocks. Most participants never asked to step out but indicated that they could “keep going”, in which case the distractor math task began immediately after the experimental block. The researcher asked the participants how they were feeling at the end of each block and, based on the observation during piloting, that most participants felt like it was difficult to remain seated for the entire duration of the study phase, after blocks 4, 5, and 6, the researcher suggested a walk outside the lab as part of the distractor math task. During the walk, participants performed mental math operations that the researcher asked them to perform. A few participants indicated that they did not feel like taking a walk – these participants performed the distractor math task in its entirety in the lab.

The six experimental blocks were followed by the recency block, which was composed of 70 trials. After the recency block, participants were given a 15-min break, during which they were free to leave the lab. Upon their return to the lab, the participants performed Posttests 1, 2, and 3, in that sequence. Participants were given unlimited time to perform these tasks. This was done to make sure that any knowledge that they had was captured and not only that which they could produce within a limited time window. This also took into account the fact that participants may differ in how quickly they perform the tasks. The immediate posttests were followed by the completion of the background questionnaire. After this, participants received cash compensation for session one.

Participants were asked to return for the second session two weeks after session one. However, not all participants were able to come back exactly two weeks after session one. For the participants who were not able to come back after two weeks, session two was mostly conducted with a shorter retention interval between the two sessions. Participants were not told anything about the content of the second session. Session two was identical in content to the immediate posttests. At the end of session two, participants were asked whether they had had any exposure to the targeted Finnish words outside of the lab between the two sessions.

This was noted by the researcher. All participants except one (whose delayed posttest data was removed from the analysis) stated that they had had no such exposure. At the end of session two, participants received cash compensation for the session.

Analyses

SPSS version 25 (IBM Corp., 2017) was used for all statistical analyses in this study. SPSS version 25, Microsoft Office 365 Excel and PowerPoint were used for data management and some of the graphics. Linear Mixed modeling and Moderated Mediation analyses were used. All statistical analyses are two-tailed and conducted at an alpha level of .05 except for cases where a Bonferroni correction is performed to adjust for multiple testing.

CHAPTER 4

RESULTS

Background questionnaire

See the *Participants* section for demographic information collected through the background questionnaire. Most participants noted that none of the words struck them as familiar. No participants were able to produce the correct translation upon initial encounter, indicating no prior knowledge. Six participants noted that some or many of the words looked like Spanish words or words from other languages in terms of the spelling.

Posttests results

To answer the first research question, which asks whether the length of the interval between repeated retrieval events and the amount of time given for study of an L2-L1 translation as feedback affect learning from retrieval practice in a PAL format, posttest results were examined as a function of ISI and study time. The no-practice condition was used as a baseline in some of the analyses to isolate more effectively the effects of retrieval practice at different levels of ISI.

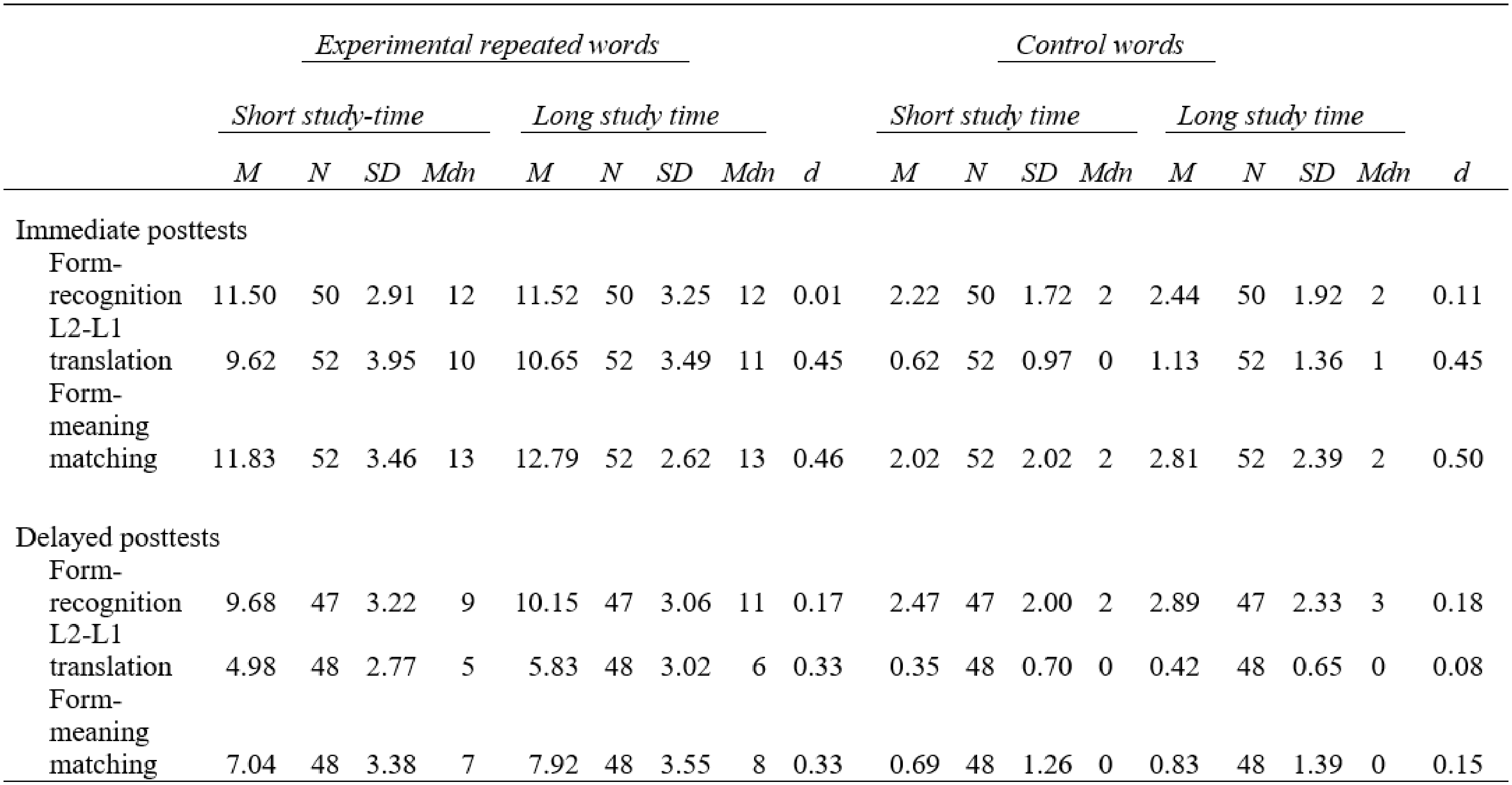
Reliability for the six posttests was as follows: immediate form-recognition test: α = .694; immediate L2-L1 translation test: Cronbach's α = .790; immediate form-meaning mapping test: Cronbach's α = .789; delayed form-recognition test: Cronbach's α = .779; delayed L2-L1 translation test: Cronbach's α = .724; delayed form-meaning mapping test: Cronbach's α = .882. Accuracy was acceptable for all participants (< 10%) except that for two participants on the immediate Posttest 1 and one participant on the delayed Posttest 1. These participants' data were excluded for the corresponding tests.

Four participants did not come back for the delayed posttest. Therefore, these participants only provided immediate posttest data. The posttests were scored as follows: one point was awarded for each correct response and zero points were awarded for an incorrect response or no response (where participants did not underline a target word on Posttest 1 or did not attempt to write its translation on Posttest 2 or did not attempt to match it with a translation on Posttest 3). Not all participants were able to come back two weeks after session one; therefore, there is a number of levels of time of delayed test in the present data. Participants can be divided in to two groups: 21 participants who came back 6-8 days after session one and 26 participants who came back 11-16 days after session one.

Posttest results: Descriptive statistics. Table 5 presents raw scores for the three immediate and delayed posttests in the experimental and control conditions separately for the shorter and longer study time duration conditions. Here, each score is out of 18 possible points (as there are 36 words in the experimental and in the control condition and half of each was presented under the short study time condition while the other half was presented under the long study time condition for any given participant). Cohen's d effect sizes were calculated relative to the results in the short study time condition to investigate the effect of study time.

The results show that there is a small effect of study time across the practice and no-practice conditions and across the two retention intervals (immediate vs. delayed test). There is further a positive effect of repetition in these numbers. Recall that in the control condition, no true retrieval attempts occurred for the target items, as here, the words were studied only once, while in the experimental condition participants were additionally given the opportunity for five true retrieval attempts and five additional restudy opportunities.

Table 5: *Raw posttest scores in the practice and no-practice conditions*



	Experimental repeated words										*Control words*								
	Short study-time				*Long study time*					*Short study time*				*Long study time*				
	M	*N*	*SD*	*Mdn*	*M*	*N*	*SD*	*Mdn*	*d*	*M*	*N*	*SD*	*Mdn*	*M*	*N*	*SD*	*Mdn*	*d*
Immediate posttests																		
Form-recognition	11.50	50	2.91	12	11.52	50	3.25	12	0.01	2.22	50	1.72	2	2.44	50	1.92	2	0.11
L2-L1 translation	9.62	52	3.95	10	10.65	52	3.49	11	0.45	0.62	52	0.97	0	1.13	52	1.36	1	0.45
Form-meaning matching	11.83	52	3.46	13	12.79	52	2.62	13	0.46	2.02	52	2.02	2	2.81	52	2.39	2	0.50
Delayed posttests																		
Form-recognition	9.68	47	3.22	9	10.15	47	3.06	11	0.17	2.47	47	2.00	2	2.89	47	2.33	3	0.18
L2-L1 translation	4.98	48	2.77	5	5.83	48	3.02	6	0.33	0.35	48	0.70	0	0.42	48	0.65	0	0.08
Form-meaning matching	7.04	48	3.38	7	7.92	48	3.55	8	0.33	0.69	48	1.26	0	0.83	48	1.39	0	0.15

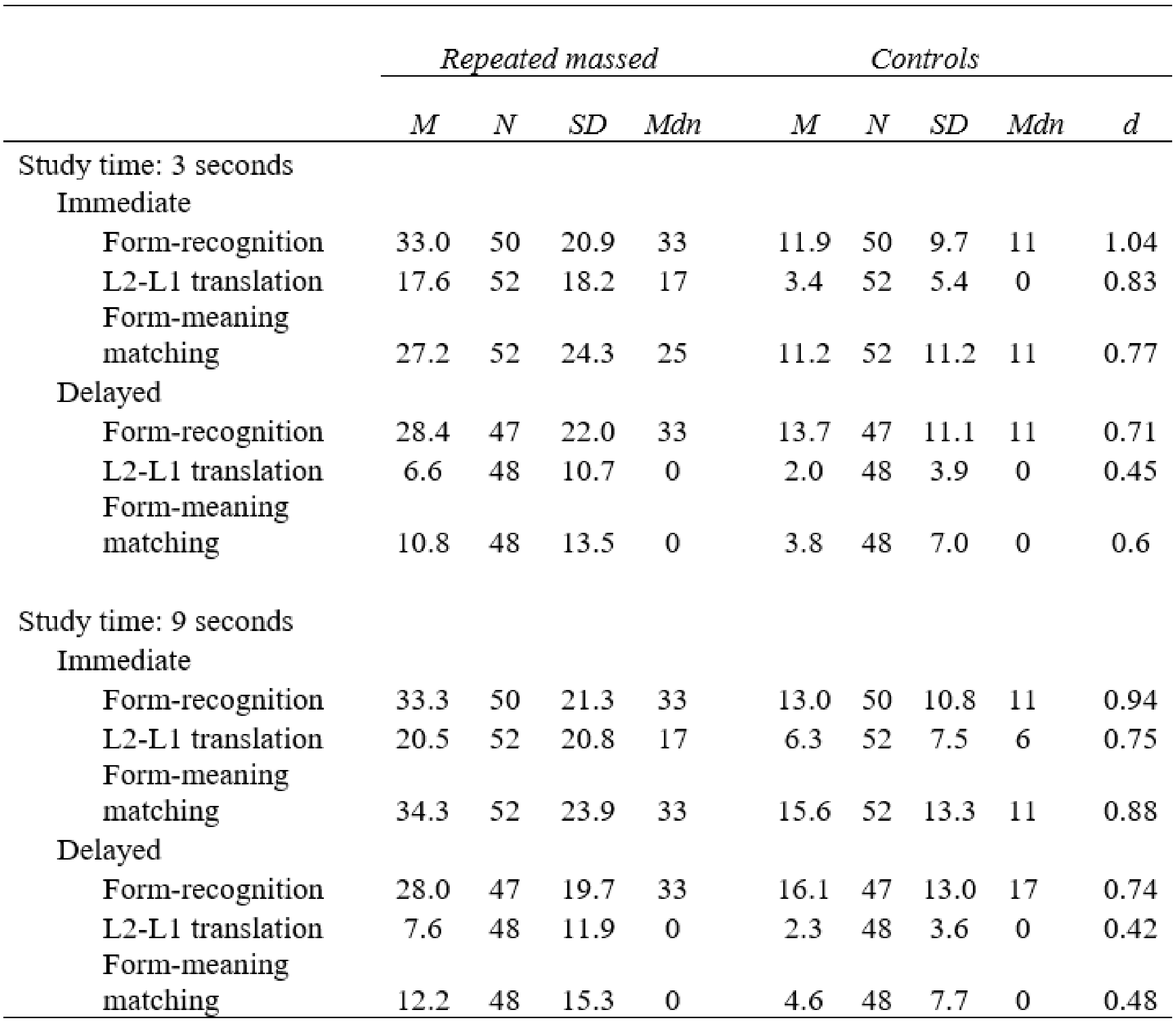
Table 6: *Raw posttest scores across the three experimental conditions*

Practice type:	*Massed*				*Short-spaced*				*Long-spaced*				*d*	*d*
	M	*N*	*SD*	*Mdn*	*M*	*N*	*SD*	*Mdn*	*M*	*N*	*SD*	*Mdn*	*(Short-spaced)*	*(Long-spaced)*
Short presentation duration														
Immediate														
Form-recognition	1.98	50	1.25	2	4.70	50	1.31	5	4.82	50	1.16	5	1.90	1.81
L2-L1 translation	1.06	52	1.09	1	4.31	52	1.71	5	4.25	52	1.78	5	2.10	2.00
Form-meaning matching	1.63	52	1.46	1	5.06	52	1.35	6	5.13	52	1.37	6	2.21	2.33
Delayed														
Form-recognition	1.70	47	1.32	2	3.98	47	1.50	4	4.00	47	1.50	4	1.42	1.40
L2-L1 translation	0.40	48	0.64	0	2.13	48	1.44	2	2.46	48	1.37	3	1.31	1.66
Form-meaning matching	0.65	48	0.81	0	3.04	48	1.71	3	3.35	48	1.38	3	1.74	2.53
Long presentation duration														
Immediate														
Form-recognition	2.00	50	1.28	2	4.76	50	1.42	5	4.76	50	1.42	5	1.83	1.83
L2-L1 translation	1.23	52	1.25	1	4.69	52	1.59	5	4.73	52	1.44	5	2.04	2.39
Form-meaning matching	2.06	52	1.43	2	5.44	52	0.94	6	5.29	52	1.14	6	2.20	2.10
Delayed														
Form-recognition	1.68	47	1.18	2	4.23	47	1.37	5	4.23	47	1.36	4	1.81	1.70
L2-L1 translation	0.46	48	0.71	0	2.38	48	1.50	2	3.00	48	1.71	3	1.17	1.53
Form-meaning matching	0.73	48	0.92	0	3.27	48	1.80	4	3.92	48	1.66	4	1.40	2.09

Table 6 presents raw scores for the three immediate and delayed posttests in the three ISI conditions separately. The scores are out of six possible points. Effect sizes are calculated relative to the scores in the massed condition to explore any benefits of spacing practice. Table 6 shows a considerable difference between the massed and the two spaced conditions across the different test types and different levels of RI. The difference between the two spaced conditions is smaller and is not consistent. There appears to be a small lag effect, whereby the longer spaced condition produced slightly better scores, particularly in the delayed posttests. The numbers also show a small benefit of longer study time that is, again, quite consistent across the conditions, test types, and RIs.

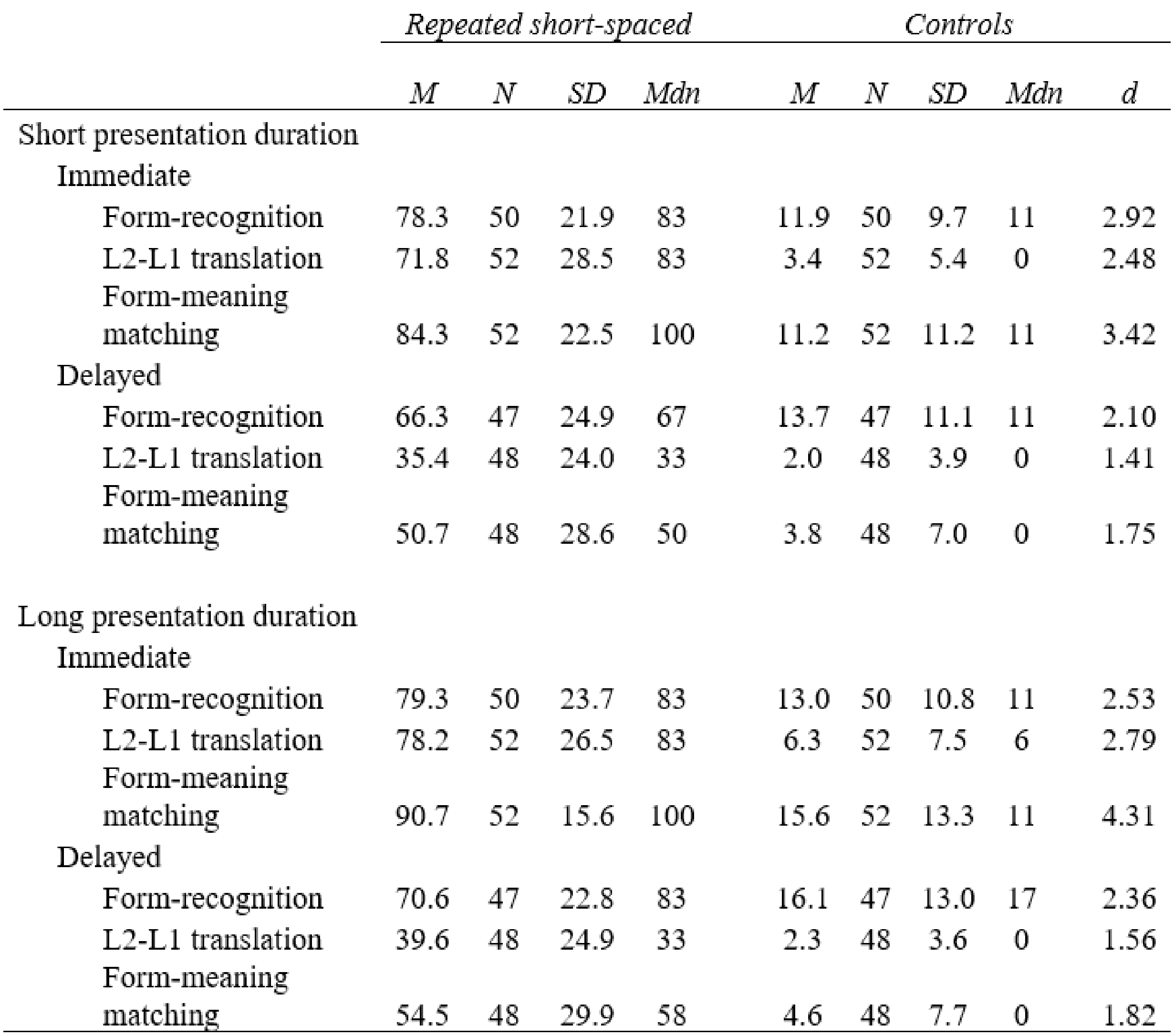
Tables 7-9 present the scores across the three ISI conditions and in the control condition as percentages. Percentages are presented because of the difference in the number of possible correct responses between the control condition and each ISI condition. The Cohen’s *d* effect sizes are calculated relative to the no-practice control condition, to investigate the effects of repetition in the three different repetition schedules.

Table 7: *Percent correct in the massed practice and no-practice conditions*



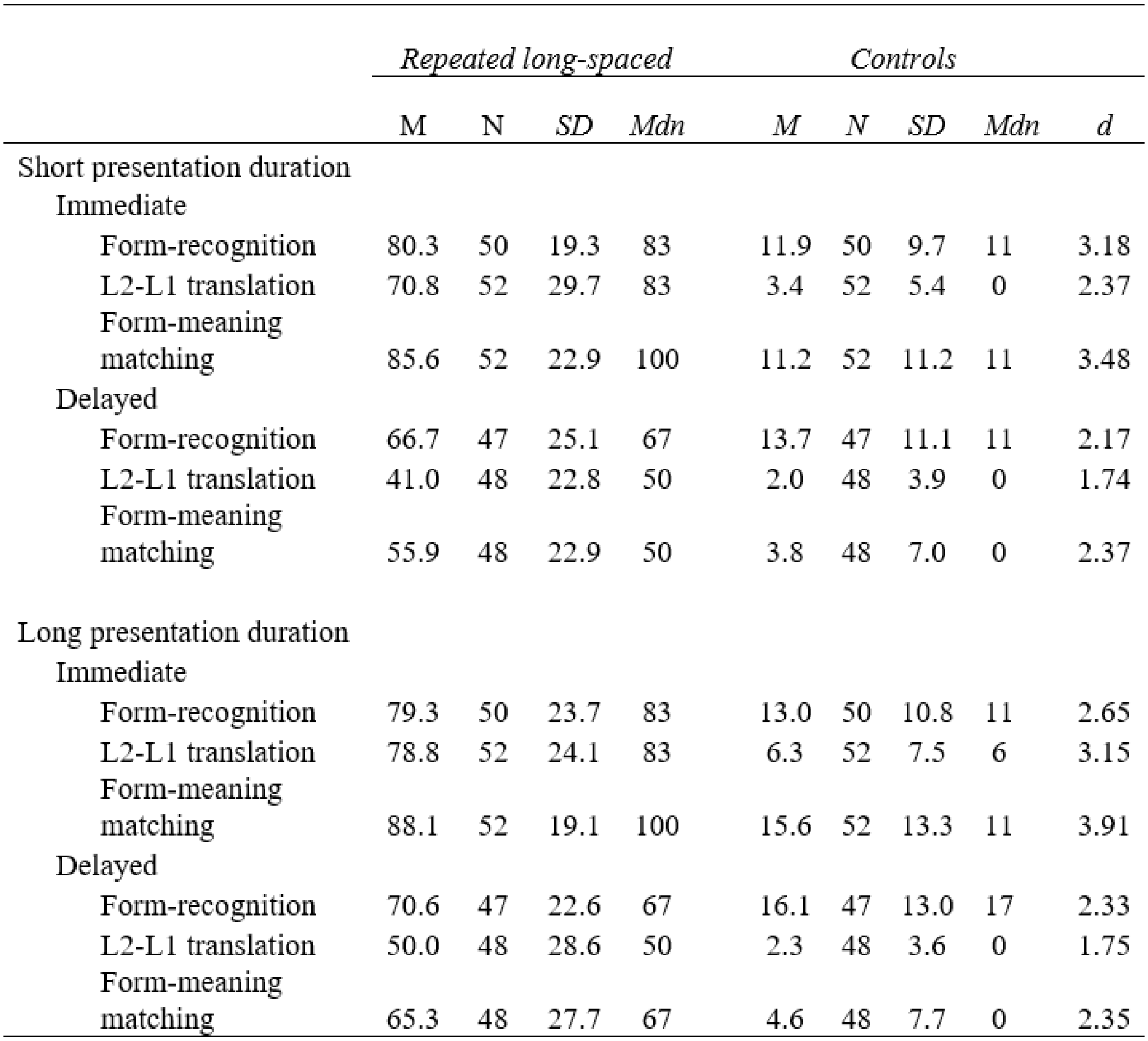
	Repeated massed				*Controls*				
	M	*N*	*SD*	*Mdn*	*M*	*N*	*SD*	*Mdn*	*d*
Study time: 3 seconds									
Immediate									
Form-recognition	33.0	50	20.9	33	11.9	50	9.7	11	1.04
L2-L1 translation	17.6	52	18.2	17	3.4	52	5.4	0	0.83
Form-meaning matching	27.2	52	24.3	25	11.2	52	11.2	11	0.77
Delayed									
Form-recognition	28.4	47	22.0	33	13.7	47	11.1	11	0.71
L2-L1 translation	6.6	48	10.7	0	2.0	48	3.9	0	0.45
Form-meaning matching	10.8	48	13.5	0	3.8	48	7.0	0	0.6
Study time: 9 seconds									
Immediate									
Form-recognition	33.3	50	21.3	33	13.0	50	10.8	11	0.94
L2-L1 translation	20.5	52	20.8	17	6.3	52	7.5	6	0.75
Form-meaning matching	34.3	52	23.9	33	15.6	52	13.3	11	0.88
Delayed									
Form-recognition	28.0	47	19.7	33	16.1	47	13.0	17	0.74
L2-L1 translation	7.6	48	11.9	0	2.3	48	3.6	0	0.42
Form-meaning matching	12.2	48	15.3	0	4.6	48	7.7	0	0.48

Table 8: *Percent correct in the short-spaced practice and no-practice conditions*



	Repeated short-spaced				*Controls*				
	M	*N*	*SD*	*Mdn*	*M*	*N*	*SD*	*Mdn*	*d*
Short presentation duration									
Immediate									
Form-recognition	78.3	50	21.9	83	11.9	50	9.7	11	2.92
L2-L1 translation	71.8	52	28.5	83	3.4	52	5.4	0	2.48
Form-meaning matching	84.3	52	22.5	100	11.2	52	11.2	11	3.42
Delayed									
Form-recognition	66.3	47	24.9	67	13.7	47	11.1	11	2.10
L2-L1 translation	35.4	48	24.0	33	2.0	48	3.9	0	1.41
Form-meaning matching	50.7	48	28.6	50	3.8	48	7.0	0	1.75
Long presentation duration									
Immediate									
Form-recognition	79.3	50	23.7	83	13.0	50	10.8	11	2.53
L2-L1 translation	78.2	52	26.5	83	6.3	52	7.5	6	2.79
Form-meaning matching	90.7	52	15.6	100	15.6	52	13.3	11	4.31
Delayed									
Form-recognition	70.6	47	22.8	83	16.1	47	13.0	17	2.36
L2-L1 translation	39.6	48	24.9	33	2.3	48	3.6	0	1.56
Form-meaning matching	54.5	48	29.9	58	4.6	48	7.7	0	1.82

Table 9: *Percent correct in the long-spaced practice and no-practice conditions*



	Repeated long-spaced				*Controls*				
	M	N	*SD*	*Mdn*	*M*	*N*	*SD*	*Mdn*	*d*
Short presentation duration									
Immediate									
Form-recognition	80.3	50	19.3	83	11.9	50	9.7	11	3.18
L2-L1 translation	70.8	52	29.7	83	3.4	52	5.4	0	2.37
Form-meaning matching	85.6	52	22.9	100	11.2	52	11.2	11	3.48
Delayed									
Form-recognition	66.7	47	25.1	67	13.7	47	11.1	11	2.17
L2-L1 translation	41.0	48	22.8	50	2.0	48	3.9	0	1.74
Form-meaning matching	55.9	48	22.9	50	3.8	48	7.0	0	2.37
Long presentation duration									
Immediate									
Form-recognition	79.3	50	23.7	83	13.0	50	10.8	11	2.65
L2-L1 translation	78.8	52	24.1	83	6.3	52	7.5	6	3.15
Form-meaning matching	88.1	52	19.1	100	15.6	52	13.3	11	3.91
Delayed									
Form-recognition	70.6	47	22.6	67	16.1	47	13.0	17	2.33
L2-L1 translation	50.0	48	28.6	50	2.3	48	3.6	0	1.75
Form-meaning matching	65.3	48	27.7	67	4.6	48	7.7	0	2.35

This comparison shows that the beneficial effect of repetition is seen across the three ISI conditions, although it is much smaller in the massed condition than in the two spaced conditions. This suggests that massed retrieval practice may have little benefit over a single study event. In fact, median values for scores on some of the tests (particularly in the delayed tests) are zero in the massed condition, suggesting no knowledge gained from massed retrieval practice. Although increasing the time a learner spends studying an L2-L1 translation pair per repetition and in total seems to benefit learning, even when this is done through simple maintenance rehearsal, spacing repeated retrieval practice appears to have a larger benefit than does increasing study duration.

Figures 4-6 present the scores on the three immediate and delayed posttests across the three ISI conditions.

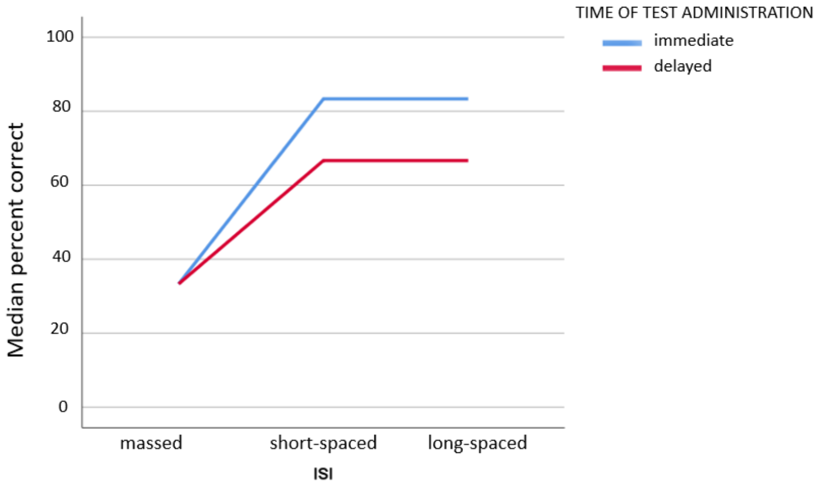


Figure 4: Form-recognition scores in the three ISI conditions

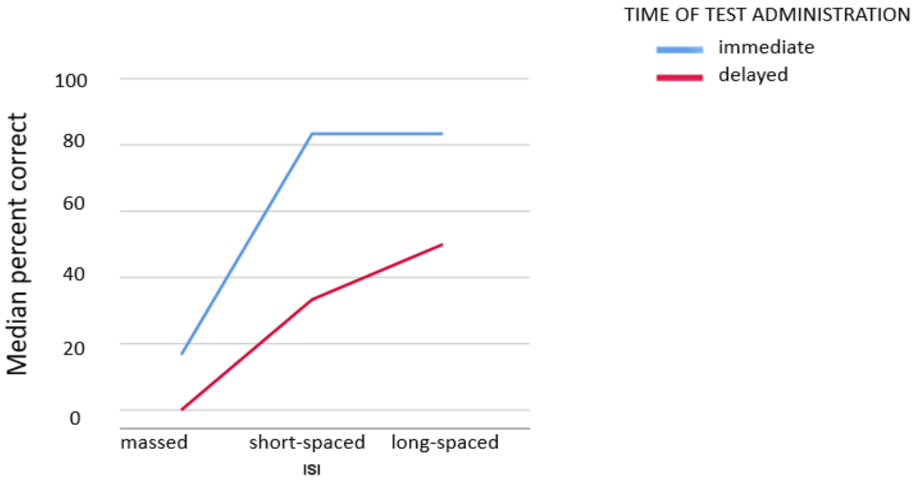


Figure 5: L2- L1 translation scores in the three ISI conditions

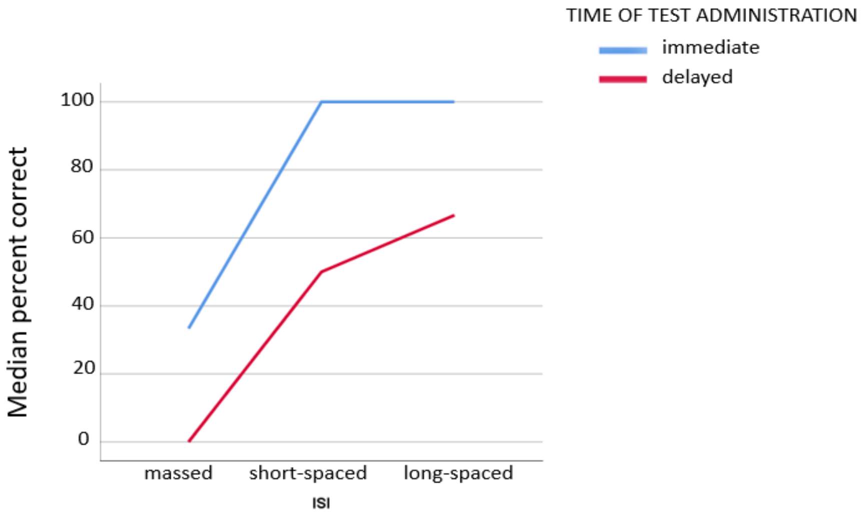


Figure 6: Form-meaning mapping scores in the three ISI conditions

For each test type, there was a considerable increase in posttest scores between the massed and the short-spaced condition both on immediate and delayed test iterations. However, the difference between the short- and long-spaced conditions seems to differ across test iterations: there seems to be no difference between the two spaced conditions on the immediate posttests, however, there seems to be an increase in the scores from the short- to the long-spaced condition in the delayed posttests. The delayed posttests all show lower scores than the scores on the immediate posttests, with the difference being relatively smaller in the form-recognition posttests. The difference between the immediate and delayed posttest scores indicates a forgetting process. The pattern of results suggests a slower rate of forgetting in the longer spaced condition than in the shorter spaced condition.

Participants differed with respect to time of delayed test. The time of delayed test will be taken into account in statistical tests. The different retention intervals center around one and two weeks. Further, based on the fact that there is a break in the continuity of RI lengths that mirrors that in forgetting slopes, differences in scores will be investigated descriptively between the resulting two groups of participants. Factor scores from a principle component analysis were used here for a more succinct presentation of scores. Table 10 presents the correlations among the three test types as well as the results of the principal component analysis.

Table 10: *Correlations and loadings for each test on the extracted component*

	Pearson correlation coefficients			Principal component analysis
Posttest	Form-recognition	L2-L1 translation	Form-meaning matching	Component 1 (89% variance explained)
Form-recognition	1			.908
L2-L1 translation	.813***	1		.963
Form-meaning matching	.801***	.926***	1	.959

****p* < .001

The three posttests load quite highly on the extracted component and the variance explained by this component alone is quite high. *Figure 7* presents the rate of forgetting in the two groups of participants that differ with respect to time of delayed test (one vs. two weeks).

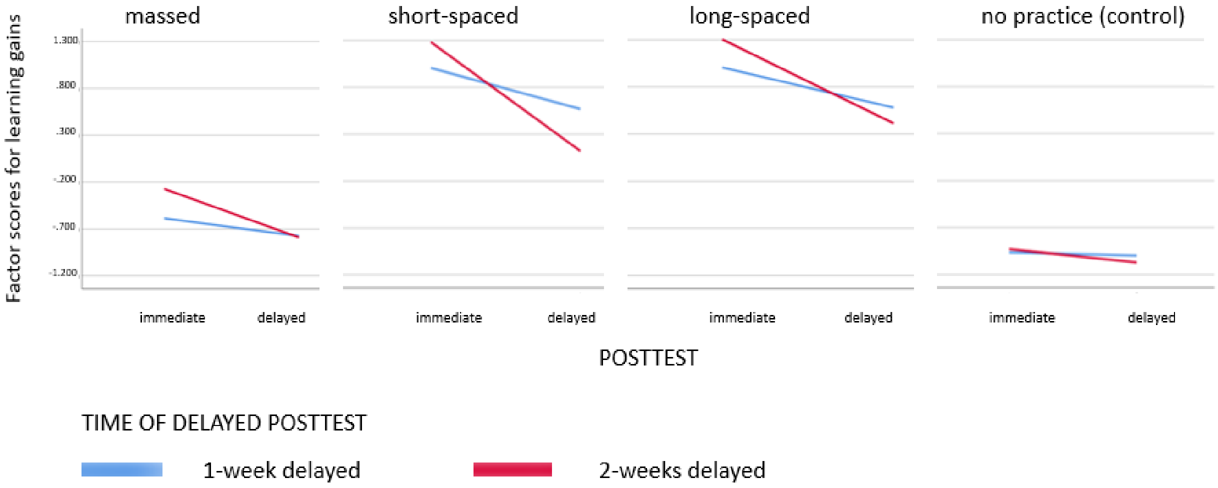


Figure 7: Posttest results in the three ISIs for the two groups of participants

Figure 7 shows that the group that returned for the delayed posttests two weeks after the study phase had a steeper forgetting slope than the group that returned one week after the study-phase. However, it also shows that the former group had higher scores on most of the immediate tests, suggesting that the two groups of participants differ with respect to knowledge gained and this difference is independent of time of delayed test administration. *Figure 8* presents these scores separately in the two study time conditions.

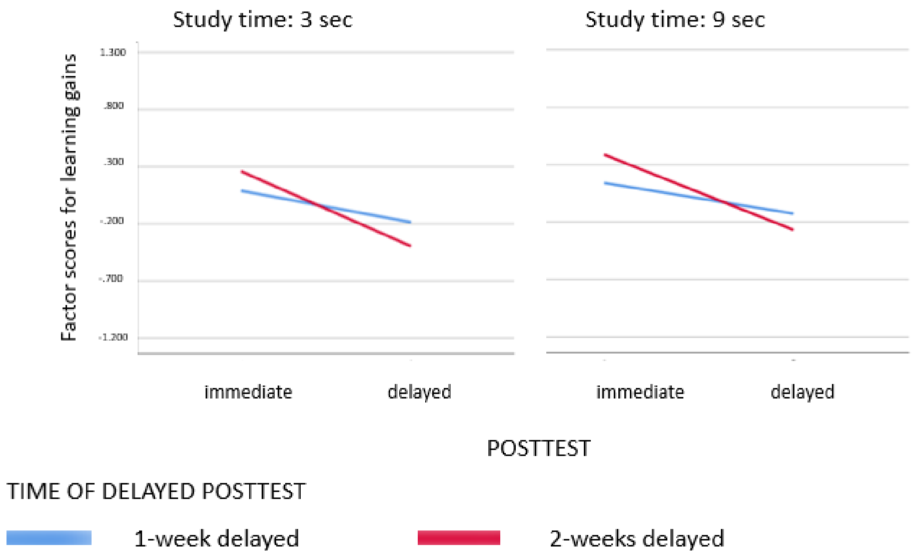


Figure 8: Effect of study time on scores in the two groups

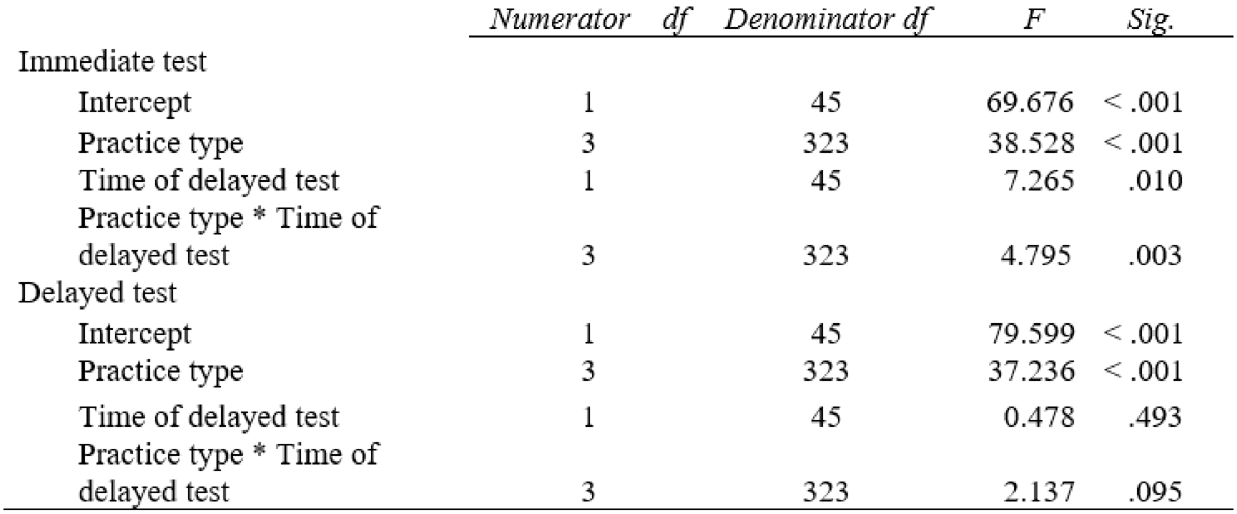
For both study duration conditions, there is a similar pattern of a steeper slope between the immediate and delayed posttests, but also higher scores on the immediate posttests, in the group that took the delayed posttests two weeks after the study phase, again suggesting a difference between the two groups that may be independent of time of delayed test.

Posttest results: Inferential statistics. An omnibus test including the immediate and delayed scores in a long format was run for each of the three test types. I included ISI, RI, and study time as the independent variables. Linear mixed modeling was used to account for the nested structure of the data, as here multiple data points were contributed by each of the participants. Because participants varied in the time between the immediate and delayed posttests, which means that they likely differed in the forgetting slopes between the two tests (the RI variable), a random slope was included for this level-two variable to control for such differences. The unstructured covariance type was selected as the most robust type. Because of the large number of independent variables, I used a simultaneous entry and Restricted Maximum Likelihood (REML) estimation. Due to high collinearity between the two variables of ISI and the variable that distinguishes experimental items from control items, these were collapsed into one variable that in these analyses will be called practice type. Thus, the practice type variable used in these analyses includes the three levels of temporal distribution of repeated encounters and one level of non-repeated control words.

The form-recognition test. The residuals for the form-recognition test were close to normally distributed with 3 outliers beyond -3SD, which were removed. The removal of the outliers resulted in a normal distribution according to the Kolmogorov-Smirnov ($p = .200$) and Shapiro-Wilk ($p = .832$) tests of normality. The distribution further had skewness and kurtosis values within acceptable ranges (skewness = -.022, $SE_{\text{skewness}} = .089$; kurtosis = -.025, $SE_{\text{kurtosis}} = .178$). For this reason, no data transformation was performed and, instead, raw percent correct scores were used. The ICC was .059, suggesting that roughly 6% of the variance in the dependent variable was attributable to the effect of participant differences. While this, again, is a small value of ICC, I used multi-level modeling because the software used allowed such an analysis but also because a random slope was of interest in the present case.

The omnibus analysis revealed a significant interaction between RI (immediate vs. delayed test) and practice type, $F_{(3,\, 669.768)} = 6.659$, $p < .001$, but no other significant interactions (all ps > .05). There was further a significant main effect of practice type, $F_{(3,\, 669.768)} = 675.566$, $p < .001$, and a main effect of RI, $F_{(1,\, 129.349)} = 8.916$, $p = .003$. Study time did not interact with any of the variables (all ps > .05) and also did not have a significant main effect, β = 2.364, $F_{(1,\ 669.768)}$ = 1.550, p = .214. To investigate the RI by practice type interaction, separate linear mixed effects analyses were run for the immediate and delayed posttests with practice type as a four-level independent variable and time of delayed test as a covariate that should affect only scores on the delayed posttest. Parameter estimates were further examined with the no-practice condition and the short-spaced condition as the reference categories in two separate analyses. This allowed to compare all the levels of practice type with a minimum number of separate comparisons. The Bonferroni correction was used to adjust the alpha level for multiple testing: α = .05/3 = .016. Table 11 presents the omnibus test results separately for the immediate and delayed test with practice type as the independent variable.

Table 11: *Form-recognition omnibus test*

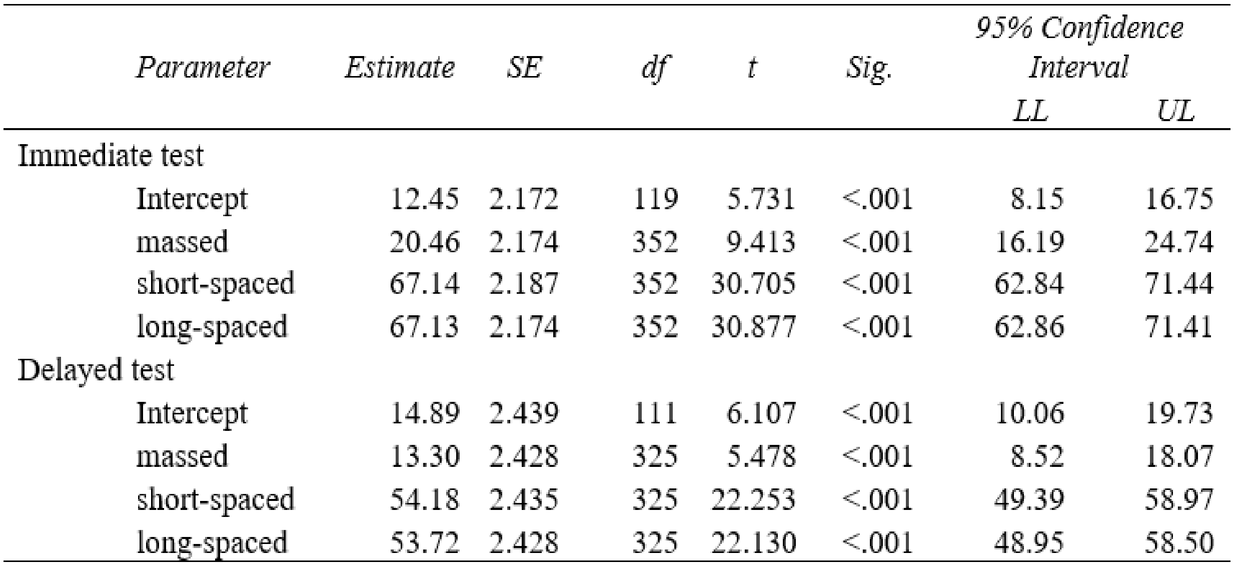


	Numerator df	*Denominator df*	*F*	*Sig.*
Immediate test				
Intercept	1	45	69.676	< .001
Practice type	3	323	38.528	< .001
Time of delayed test	1	45	7.265	.010
Practice type * Time of delayed test	3	323	4.795	.003
Delayed test				
Intercept	1	45	79.599	< .001
Practice type	3	323	37.236	< .001
Time of delayed test	1	45	0.478	.493
Practice type * Time of delayed test	3	323	2.137	.095

The results of the separate omnibus tests for the immediate and delayed posttests show a significant effect of practice type for both RIs. Further, the results show that there is actually a significant difference between the two groups of participants that differ with respect to time of delayed posttest in the immediate scores but not in the delayed scores, contrary to what should be observed. Further, this variable also interacts with practice type in the immediate scores. Time of delayed test should not have an effect on the immediate scores and should not interact with other variables in these scores, as participants in the two groups do not differ with respect to time of the immediate test. This pattern of results confirms statistically the observation from *Figure 7* that the two groups differ in their learning gains overall and that this difference may exist independently of when the delayed test is administered. For this reason, any difference between the delayed posttest scores in the two groups needs to be interpreted with caution.

Table 12 presents parameter estimates for a comparison between the effect of practice under the three practice type conditions against the no-practice condition on the immediate and delayed form-recognition tests. Here, the estimates are all in raw percentages. The intercept respresents the mean score in the no-practice condition and each slope represents the mean difference between the no-practice condition and the corresponding practice schedule condition. The null hypothesis for the effect of intercept is that the mean of the scores in the no-practice condition is equal to zero. The null hypothesis for each slope is that the scores in the corresponding condition are not different from the scores in the no-practice condition, which is the reference category represented by the intercept.

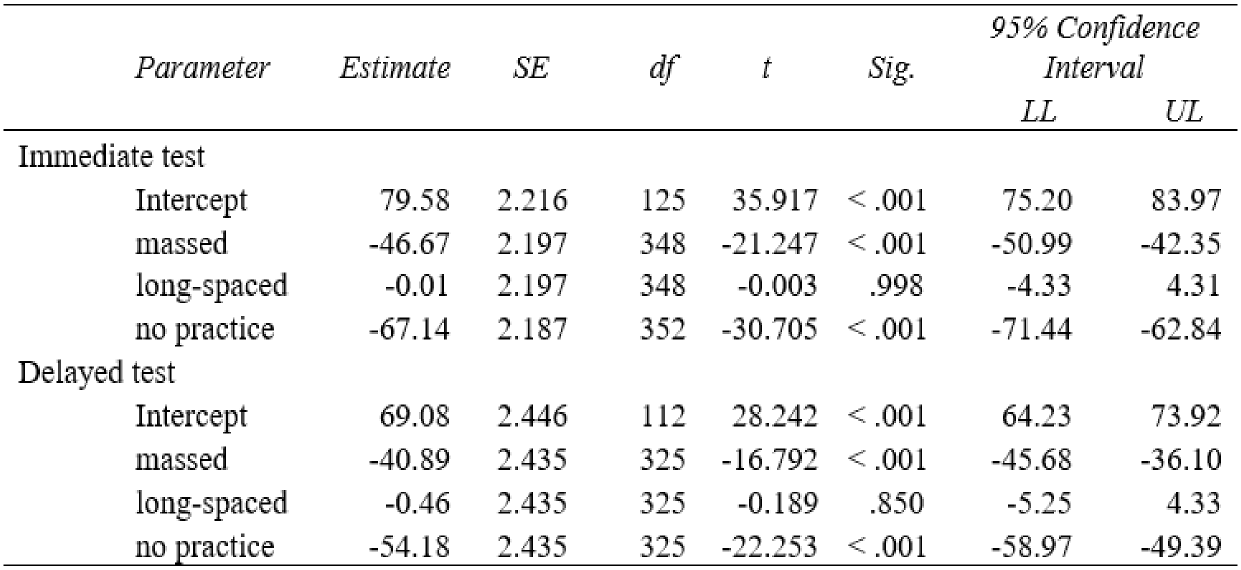
Table 12: *Form-recognition results against the no-practice condition*



Parameter	Estimate	SE	df	t	Sig.	95% Confidence Interval	
						LL	UL
Immediate test							
Intercept	12.45	2.172	119	5.731	<.001	8.15	16.75
massed	20.46	2.174	352	9.413	<.001	16.19	24.74
short-spaced	67.14	2.187	352	30.705	<.001	62.84	71.44
long-spaced	67.13	2.174	352	30.877	<.001	62.86	71.41
Delayed test							
Intercept	14.89	2.439	111	6.107	<.001	10.06	19.73
massed	13.30	2.428	325	5.478	<.001	8.52	18.07
short-spaced	54.18	2.435	325	22.253	<.001	49.39	58.97
long-spaced	53.72	2.428	325	22.130	<.001	48.95	58.50

The table shows that there was a significant difference between results in the no-practice condition and results in each of the practice type conditions on both immediate and delayed form-recognition tests. Further, the slopes are positive throughout, indicating that practice under each of the temporal distributions was significantly better than no retrieval practice at all and this is true of whether the learning gains are measured 30 minutes or a week or two after the study phase. However, the slopes are of different magnitudes. Thus, the effect of retrieval practice in the massed condition is considerably smaller than that in the two spaced conditions. This pattern holds for both immediate and delayed form-recognition tests. Table 13 presents parameter estimates for a comparison between scores in the short-spaced practice and the other practice schedules as well as the no-practice condition.

Table 13: *Form-recognition results against the short-spaced practice condition*



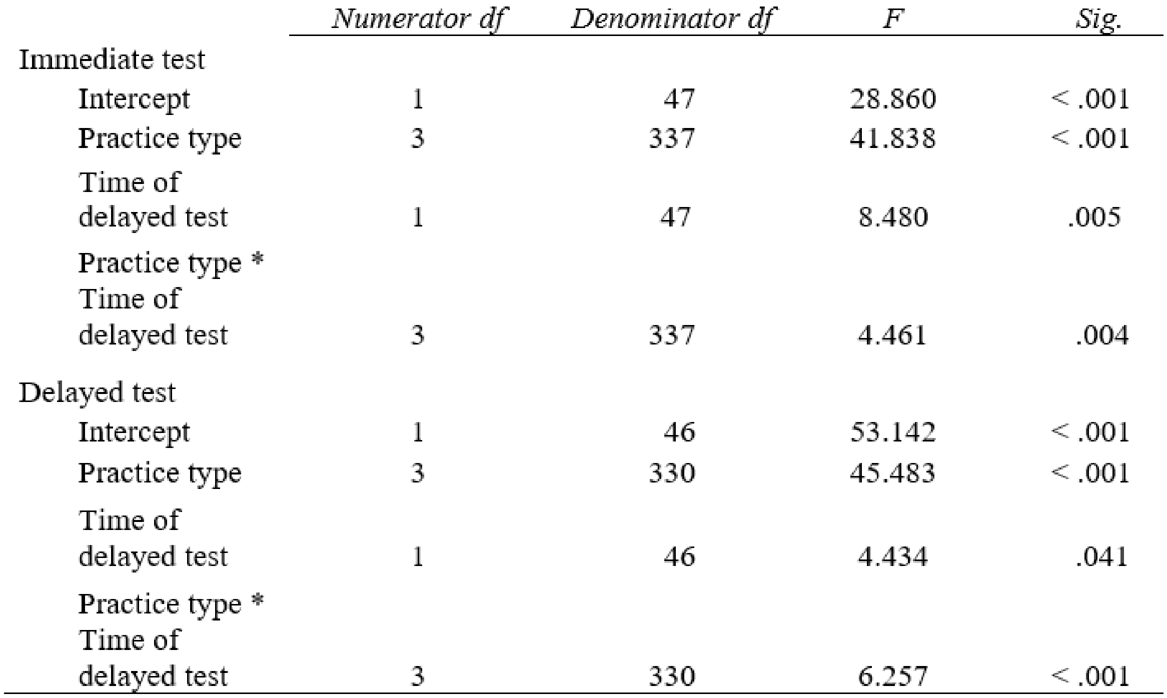
	Parameter	*Estimate*	*SE*	*df*	*t*	*Sig.*	*95% Confidence Interval*	
							LL	*UL*
Immediate test								
	Intercept	79.58	2.216	125	35.917	< .001	75.20	83.97
	massed	-46.67	2.197	348	-21.247	< .001	-50.99	-42.35
	long-spaced	-0.01	2.197	348	-0.003	.998	-4.33	4.31
	no practice	-67.14	2.187	352	-30.705	< .001	-71.44	-62.84
Delayed test								
	Intercept	69.08	2.446	112	28.242	< .001	64.23	73.92
	massed	-40.89	2.435	325	-16.792	< .001	-45.68	-36.10
	long-spaced	-0.46	2.435	325	-0.189	.850	-5.25	4.33
	no practice	-54.18	2.435	325	-22.253	< .001	-58.97	-49.39

In both the immediate and delayed form-recognition tests, the massed retrieval practice schedule produced significantly lower scores than did the short-spaced retrieval practice schedule. In both tests, there was no significant difference between the long-spaced retrieval practice schedule and the short-spaced retrieval practice schedule, with the former showing a very small nonsignificant negative slope relative to the latter, indicating a nonsignificant nonmonotonic function of lag. Further, as expected based on the previous comparisons, where the massed retrieval practice schedule was shown to produce higher scores than no practice, the no-practice condition produced significantly lower scores than the short-spaced retrieval practice schedule.

The L2-L1 translation test. In the L2-L1 translation test, participants did not need to select target forms but were presented with them and were asked to recall their meanings. The residuals for this test were close to normally distributed with four outliers in the lower tail (2 from the short-spaced and 2 from the long-spaced condition). After the removal of these outliers, the distribution was normal according to the Kolmogorov-Smirnov (*p* = .200) and Shapiro-Wilk (*p* = .606) tests of normality. The distribution further had skewness and kurtosis values within acceptable ranges (skewness = -.116, SE_{skewness} = .088; kurtosis = -.082, SE_{kurtosis} = .176). For this reason, no data transformation was performed and, instead, raw percent correct scores were used. The ICC was .059, suggesting that roughly 6% of the variance in the dependent variable was attributable to the effect of participant. While this, again, is a small value of ICC, I used multi-level modeling because the software used allowed such an analysis but also because a random slope was of interest in the present case.

The same independent variables were used as those in the form-recognition test presented above. The omnibus analysis revealed a significant interaction between RI and practice schedule, $F_{(3,\ 690.080)} = 47.297$, $p < .001$, but no other significant interactions (all *p*s > .05), as in the results of the form-recognition test. Similarly to the results of the form-recognition test presented above, there was also a significant main effect of practice type, $F_{(3,\ 690.080)} = 636.334$, $p < .001$ and a main effect of RI, $F_{(1,\ 129.702)} = 95.307$, $p < .001$. Unlike the results of the form-recognition test, however, there was further a main effect of study time, $\beta = 0.347$, $F_{(1,\ 690.080)} = 14.176$, $p < .001$. To investigate the RI by practice type interaction, separate linear mixed effects analyses were conducted for the immediate and delayed posttests with practice type as a four-level independent variable and time of delayed test as a covariate that should affect only scores on the delayed posttest. Parameter estimates were further examined with the no-practice condition and the short-spaced condition as the reference categories in two separate analyses. This allowed to compare all the levels of practice type with a minimum number of separate comparisons. The Bonferroni correction was used to adjust the alpha level for multiple testing: $\alpha = .05/3 = .016$. Table 14 presents the results of this analysis.

Table 14: *L2-L1 translation omnibus test*

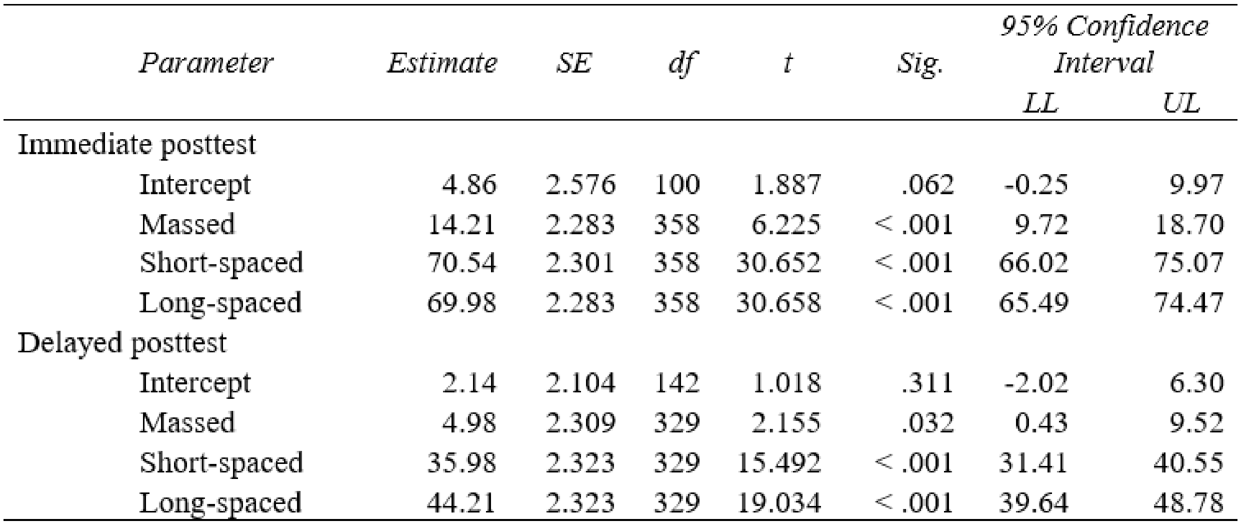


	Numerator df	*Denominator df*	*F*	*Sig.*
Immediate test				
Intercept	1	47	28.860	< .001
Practice type	3	337	41.838	< .001
Time of delayed test	1	47	8.480	.005
Practice type * Time of delayed test	3	337	4.461	.004
Delayed test				
Intercept	1	46	53.142	< .001
Practice type	3	330	45.483	< .001
Time of delayed test	1	46	4.434	.041
Practice type * Time of delayed test	3	330	6.257	< .001

There is a significant effect of practice type for both test iterations, indicating that there is a significant difference between at least two of the levels of practice in each of the two L2-L1 translation tests. Time of delayed test, the covariate, is significant for both the immediate test and the delayed test and it interacts, in both test iterations, with practice type. Because this variable cannot have an effect on the immediate scores, this again suggests that the two groups of participants that differ with regards to time of delayed test also differ in the level of knowledge gained, which in turn means that no firm conclusions can be made about the effect of time of delayed test on forgetting curves in the present case.

Table 15 presents a comparison of the different practice schedules to the no-practice condition in terms of the immediate and delayed L2-L1 translation posttest scores in percent correct translations. Here, again, the estimates are all in raw percentages and the intercept respresents the scores in the no-practice condition while each slope represents the difference between the no-practice condition and the corresponding practice schedule condition. The null hypothesis for the effect of intercept is that the scores in the no-practice condition are equal to zero. The null hypothesis for each slope is that the scores in the corresponding condition are not different from the scores in the no-practice condition, which is the reference category represented by the intercept.

Table 15: *L2-L1 translation results against the no-practice condition*

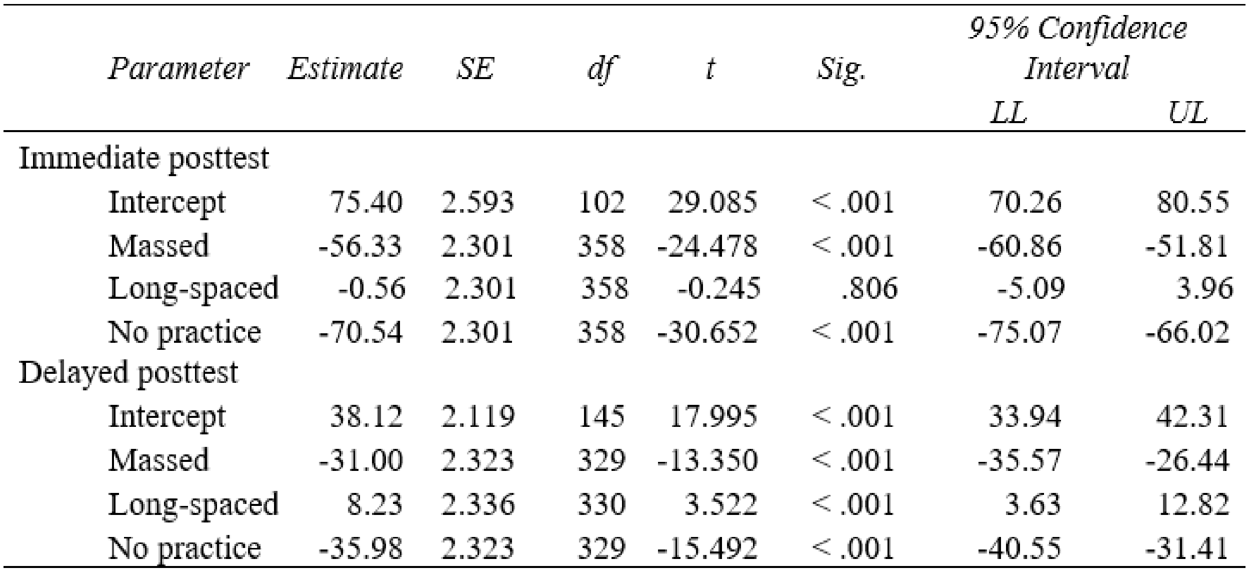


Parameter	*Estimate*	*SE*	*df*	*t*	*Sig.*	*95% Confidence Interval*	
						LL	*UL*
Immediate posttest							
Intercept	4.86	2.576	100	1.887	.062	-0.25	9.97
Massed	14.21	2.283	358	6.225	< .001	9.72	18.70
Short-spaced	70.54	2.301	358	30.652	< .001	66.02	75.07
Long-spaced	69.98	2.283	358	30.658	< .001	65.49	74.47
Delayed posttest							
Intercept	2.14	2.104	142	1.018	.311	-2.02	6.30
Massed	4.98	2.309	329	2.155	.032	0.43	9.52
Short-spaced	35.98	2.323	329	15.492	< .001	31.41	40.55
Long-spaced	44.21	2.323	329	19.034	< .001	39.64	48.78

The results show that all practice schedules resulted in significantly higher scores relative to the no-practice condition on the immediate test, although the size of the benefit varied across the different practice schedules. On the delayed test, however, there was no significant difference between the massed retrieval practice condition and the no-practice condition at the corrected alpha level, while the significant benefits of the two spaced conditions persisted across time. The nonsignificant p-value associated with the score in the no-practice condition on both the immediate and the delayed tests in turn suggests that in this condition the learning gains were close to zero.

Table 16 presents a comparison of the scores in the short-spaced retrieval practice condition against the scores in the other conditions, including the no-practice condition. The intercept respresents the scores in the short-spaced practice condition and each slope represents the difference between this condition and the corresponding practice schedule condition or the no-practice condition. The null hypothesis for the effect of intercept is that the scores in the short-spaced practice condition are equal to zero. The null hypothesis for each slope is that the scores in the corresponding condition are not different from the scores in the short-spaced practice condition, which is the reference category represented by the intercept.

Table 16: *L2-L1 translation results against the short-spaced practice condition*



	Parameter	*Estimate*	*SE*	*df*	*t*	*Sig.*	*95% Confidence Interval*	
							LL	*UL*
Immediate posttest								
	Intercept	75.40	2.593	102	29.085	< .001	70.26	80.55
	Massed	-56.33	2.301	358	-24.478	< .001	-60.86	-51.81
	Long-spaced	-0.56	2.301	358	-0.245	.806	-5.09	3.96
	No practice	-70.54	2.301	358	-30.652	< .001	-75.07	-66.02
Delayed posttest								
	Intercept	38.12	2.119	145	17.995	< .001	33.94	42.31
	Massed	-31.00	2.323	329	-13.350	< .001	-35.57	-26.44
	Long-spaced	8.23	2.336	330	3.522	< .001	3.63	12.82
	No practice	-35.98	2.323	329	-15.492	< .001	-40.55	-31.41

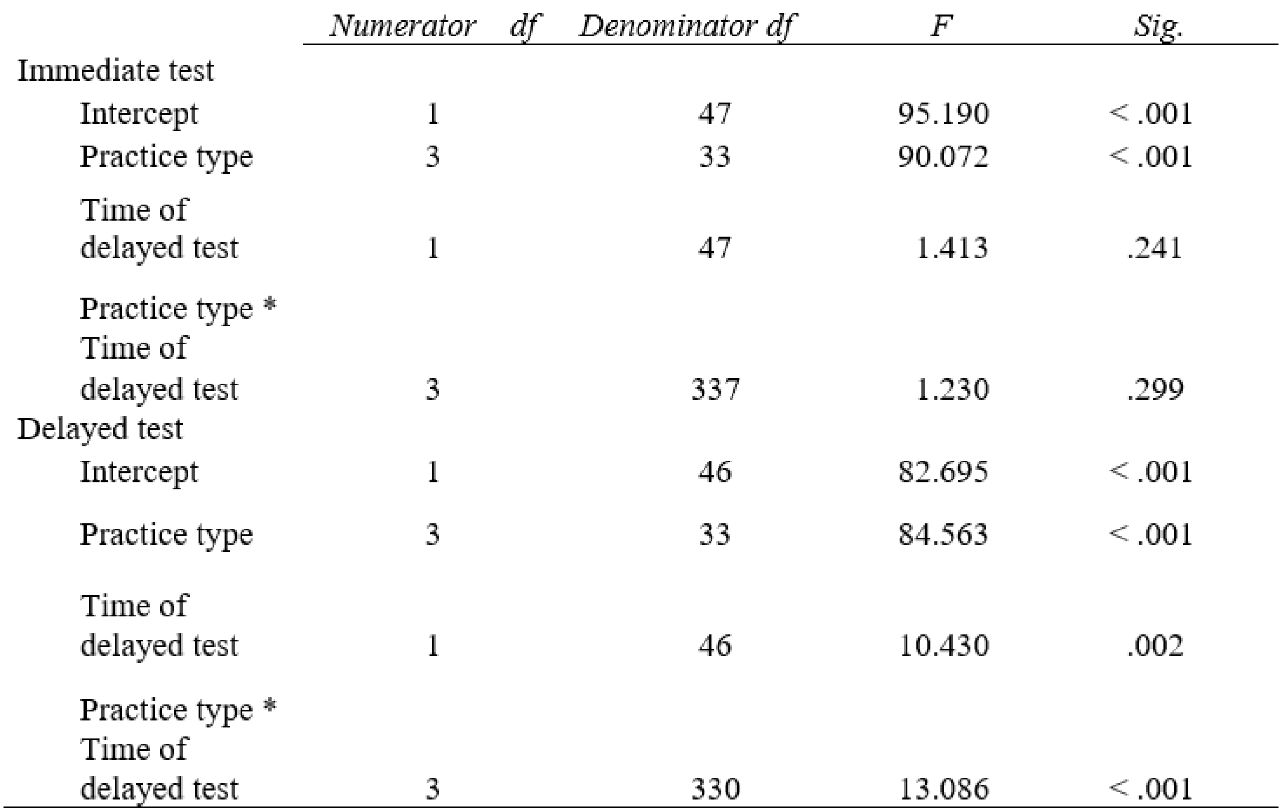
The scores in the massed condition are significantly lower than in the short-spaced condition, across the two posttests, indicating a spacing effect between these two conditions. The scores in the long-spaced condition are a tiny bit lower on the immediate posttest (showing a nonmonotonic function) though this difference is not statistically significant. However, the scores on the delayed posttest are 8% higher in the long-spaced condition than in the short-spaced condition and this difference is statistically significant, indicating a significant lag effect in the delayed L2-L1 translation posttest scores.

The form-meaning matching test. On the form-meaning matching tests, participants were presented with the Finnish words and their translations and asked to match between the two lists. The distribution of the residuals for the form-meaning matching test scores was close to normally distributed with one outlier above 3SD in the upper and one in the lower tails. These outliers were removed, which resulted in more nearly normal distribution (skewness = -.246, SE_{skewness} = .087; kurtosis = .154, SE_{kurtosis} = .174). Although the Kolmogorov-Smirnov and Shapiro-Wilk tests of normality were significant at the .05 alpha level, (p = .037 and .005, respectively), no data transformation was performed due to the fact that a .001 alpha level is recommended for these tests of normality because of how conservative they are and how sensitive they are (Field, 2013). Further, the distribution looked symmetrical and bell-shaped and the Normal Q-Q plot also did not show much deviation from the diagonal. The ICC was .059, suggesting that roughly 6% of the variance in the outcome was attributable to the effect of participant. While this, again, is a small value of ICC, I used multi-level modeling because the software used allowed such an analysis but also because a random slope was of interest in the present case.

The omnibus test showed a significant interaction between RI and practice type, $F_{(3, 691.462)}$ = 24.333, p < .001 but no other significant interactions (all ps > .05). There was further a main effect of practice type, $F_{(3, 691.462)}$ = 910.132, p < .001 and a main effect of RI, $F_{(1, 127.754)}$ = 112.564, p < .001, as well as a significant main effect of study time, β = .810, $F_{(1, 691.462)}$ = 16.456, p < .001.

To investigate the RI by practice type interaction, separate linear mixed effects analyses were conducted for the immediate and delayed posttests with practice type as a four-level independent variable and time of delayed test as a covariate that should affect only scores on the delayed posttest. Parameter estimates were further examined with the no-practice condition and the short-spaced condition as the reference categories in two separate analyses. This allowed to compare all the levels of practice type with a minimum number of separate comparisons. The Bonferroni correction was used to adjust the alpha level for multiple testing: α = .05/3 = .016. Table 17 presents the results of this analysis.

Table 17: *Form-meaning matching omnibus test*

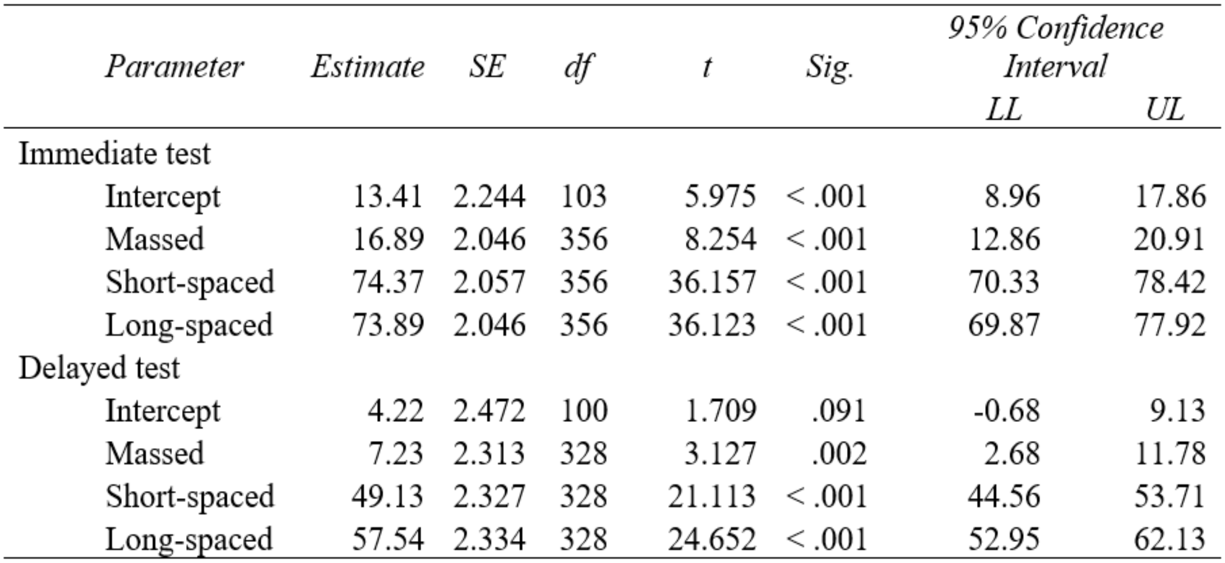


	Numerator df	*Denominator df*	*F*	*Sig.*
Immediate test				
Intercept	1	47	95.190	< .001
Practice type	3	33	90.072	< .001
Time of delayed test	1	47	1.413	.241
Practice type * Time of delayed test	3	337	1.230	.299
Delayed test				
Intercept	1	46	82.695	< .001
Practice type	3	33	84.563	< .001
Time of delayed test	1	46	10.430	.002
Practice type * Time of delayed test	3	330	13.086	< .001

The omnibus tests show that there is a significant difference between at least two of the four practice types on both the immediate and the delayed test. Further, contrary to the results of the previous two tests, here, as would logically be expected, the time of delayed test is significant only for the delayed test scores and it further significantly interacts with practice type in this test.

Table 18 presents a comparison of the different practice schedules to the no-practice condition in terms of the immediate and delayed L2-L1 form-meaning matching posttest scores in percent correct matches. Here, again, the estimates are all in raw percentages and the intercept respresents the scores in the no-practice condition while each slope represents the difference between the no-practice condition and the corresponding practice schedule condition. The null hypothesis for the effect of intercept is that the scores in the no-practice condition are equal to zero. The null hypothesis for each slope is that the scores in the corresponding condition are not different from the scores in the no-practice condition, which is the reference category represented by the intercept.

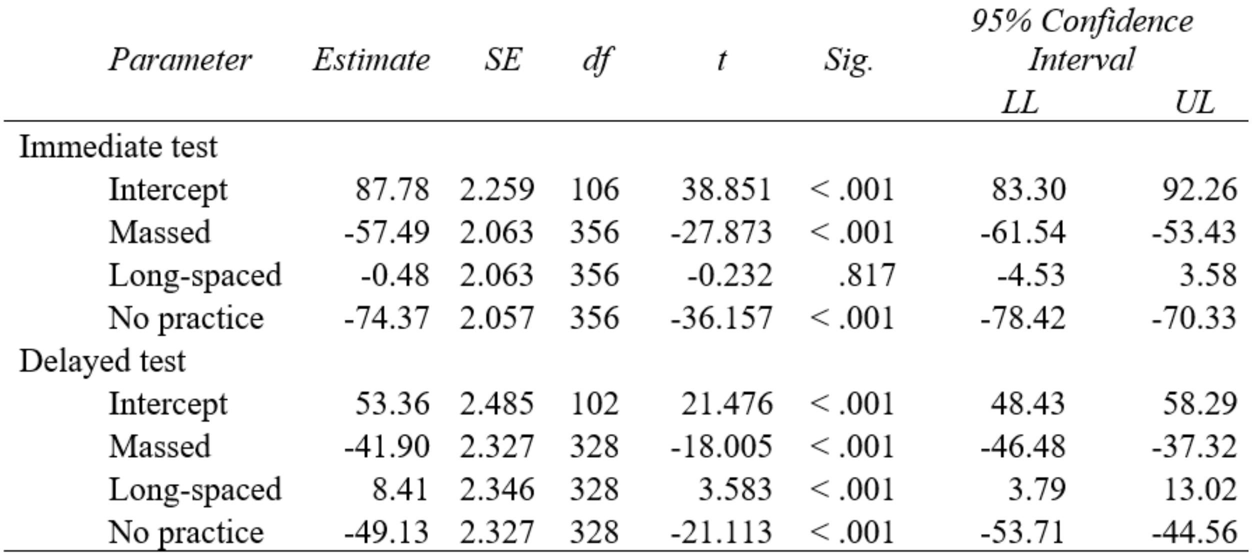
Table 18: *Form-meaning matching results against the no-practice condition*



Parameter	*Estimate*	*SE*	*df*	*t*	*Sig.*	*95% Confidence Interval*	
						LL	*UL*
Immediate test							
Intercept	13.41	2.244	103	5.975	< .001	8.96	17.86
Massed	16.89	2.046	356	8.254	< .001	12.86	20.91
Short-spaced	74.37	2.057	356	36.157	< .001	70.33	78.42
Long-spaced	73.89	2.046	356	36.123	< .001	69.87	77.92
Delayed test							
Intercept	4.22	2.472	100	1.709	.091	-0.68	9.13
Massed	7.23	2.313	328	3.127	.002	2.68	11.78
Short-spaced	49.13	2.327	328	21.113	< .001	44.56	53.71
Long-spaced	57.54	2.334	328	24.652	< .001	52.95	62.13

In the immediate and delayed posttest scores, all three practice conditions show significantly higher scores than the no-practice condition. In the delayed posttest scores, however, the benefit of massed practice over no practice is much smaller in magnitude (only 7%) while the benefits of the two spaced practice conditions remain quite large across time (49% and 58%). Table 19 presents a comparison between the scores in the short-spaced retrieval practice condition and the scores in all the other conditions, including the no-practice condition.

Table 19: *Form-meaning matching results against the short-spaced practice condition*



Parameter	*Estimate*	*SE*	*df*	*t*	*Sig.*	*95% Confidence Interval*	
						LL	*UL*
Immediate test							
Intercept	87.78	2.259	106	38.851	< .001	83.30	92.26
Massed	-57.49	2.063	356	-27.873	< .001	-61.54	-53.43
Long-spaced	-0.48	2.063	356	-0.232	.817	-4.53	3.58
No practice	-74.37	2.057	356	-36.157	< .001	-78.42	-70.33
Delayed test							
Intercept	53.36	2.485	102	21.476	< .001	48.43	58.29
Massed	-41.90	2.327	328	-18.005	< .001	-46.48	-37.32
Long-spaced	8.41	2.346	328	3.583	< .001	3.79	13.02
No practice	-49.13	2.327	328	-21.113	< .001	-53.71	-44.56

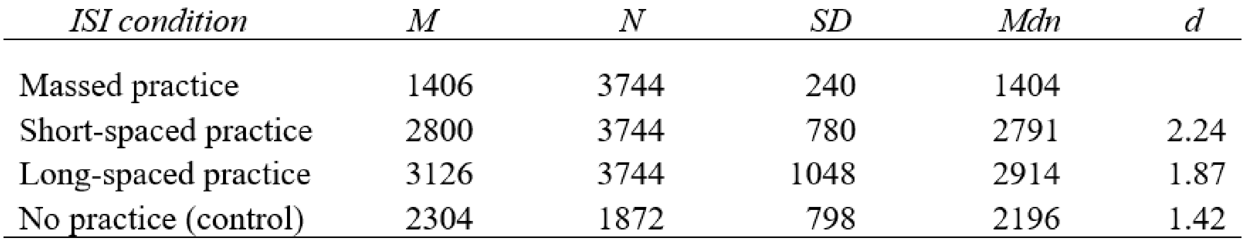
There is a significant negative slope for the massed practice and no-practice conditions relative to the short-spaced practice condition on both the immediate and the delayed posttests. There is further a small negative slope for the long-spaced practice condition relative to the short-spaced practice condition in the immediate scores (a nonmonotonic function), though this effect is not statistically significant. On the delayed posttest, by contrast, the long-spaced practice condition shows a significant 8% benefit over the short-spaced practice condition, indicating a significant lag effect. Thus, the pattern of the effect of lag is reversed between the immediate and delayed posttests. This, again, confirms the previous observation that the benefit of long-spaced practice seems to be more evident after more time, suggesting that long-spaced retrieval practice interferes more strongly with forgetting in the longer term than short-spaced or massed retrieval practice.

Study-phase results

The study phase produced quite a low percentage of errors (*M* = 2.4%, *SD* = 1.9%, Median = 1.8%, Min = 0.2%, Max = 8.8%). A cutoff point of 10% error rate was used because of the low chance of providing the correct translation for a target word by mistake due to the large number of potential translations. Errors in this case were incorrect translations given by a participant in response to either a Finnish word they had not seen before (distractors) or one they had seen before. Therefore, all participants' data were included in the study-phase analysis. Further, zero correct translations were given upon the first encounters with all the Finnish words, before a learner was given a chance to study the word, further confirming no prior knowledge of the target words. Effort indices were investigated in the three ISI conditions and also in the two study time conditions.

Study-phase response latencies: Descriptive statistics. Table 20 presents the descriptive statistics for the study-phase response latencies across the four conditions. Recall that response latencies indicate the amount of time, in milliseconds, between the moment a Finnish word appears on the screen and the time when the participant either supplies its translation by saying it aloud or states aloud that they don't know the translation. Recall, also, that the words in the experimental conditions repeated six times, while the controls were only presented once. Thus, the latencies for the control words indicate how much time or effort was spent on identifying a given word as one that had not been seen before or one for which a translation had not been seen before, or a vain search of one's memory for a nonexistent representation not encoded on any level, as the translations for these words had not been presented yet. Effect sizes in the practice conditions were calculated relative to the massed condition in order to investigate the question of whether increasing ISI leads to greater effort.

Table 20: *Response latencies across the practice conditions*

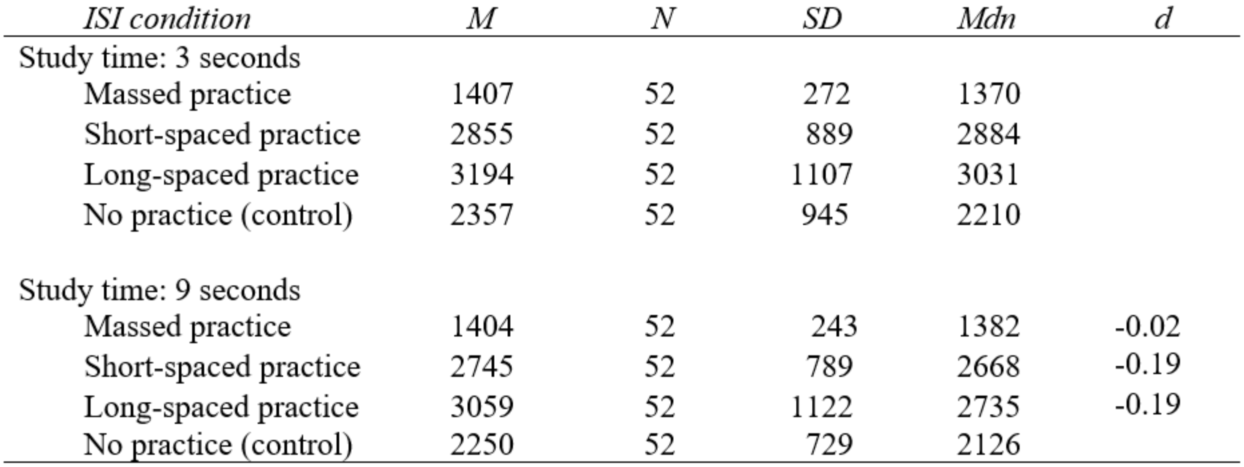


ISI condition	*M*	*N*	*SD*	*Mdn*	*d*
Massed practice	1406	3744	240	1404	
Short-spaced practice	2800	3744	780	2791	2.24
Long-spaced practice	3126	3744	1048	2914	1.87
No practice (control)	2304	1872	798	2196	1.42

On average, the least retrieval effort was observed in the massed condition. The short-spaced repetitions produced almost twice as much effort as the massed repetitions and the long-spaced repetitions produced more effort than the short-spaced repetitions though this difference is not as dramatic as that between the massed and the short-spaced repetitions. The reason for such a small difference is likely more frequent failure to recognize long-spaced repetitions as words for which a participant is able to produce the translation, which resulted in more quick "I don't know" responses without an attempt at retrieval.

Table 21 presents these statistics separately for the long and short study time, or presentation duration, conditions. Effect sizes here are calculated relative to the massed practice condition. No effect sizes are shown for the control condition because here presentation duration cannot impact response latencies as no words in this condition repeated.

Table 21: *Response latencies in the two study time conditions*



ISI condition	*M*	*N*	*SD*	*Mdn*	*d*
Study time: 3 seconds					
Massed practice	1407	52	272	1370	
Short-spaced practice	2855	52	889	2884	
Long-spaced practice	3194	52	1107	3031	
No practice (control)	2357	52	945	2210	
Study time: 9 seconds					
Massed practice	1404	52	243	1382	-0.02
Short-spaced practice	2745	52	789	2668	-0.19
Long-spaced practice	3059	52	1122	2735	-0.19
No practice (control)	2250	52	729	2126	

This table shows that words that were presented for study with their translations for 9 seconds received slightly less overall translation effort than did words that were presented for study only 3 seconds. However, this difference is very small. Further, the fact that the control condition appears to show the same pattern suggests that the difference is so small as to be easily obtained by chance. Statistical tests may help to adjudicate between these possibilities.

Table 22 presents the response latencies separately for successful and unsuccessful retrieval attempts. The first encounters are excluded from these statistics for a pure effect of success/failure, where retrieval attempts actually represented a search of one's memory for an existing memory trace. Note that the number of cases is different between these two conditions. This is because these were not set a priori by the researcher but rather were a function of participants' ability to recall a given translation and were thus outside of direct experimental control. The effect sizes here were calculated relative to the massed condition.

Table 22: *Response latencies in successful and unsuccessful retrieval attempts*

ISI condition	*M*	*N*	*SD*	*Mdn*	*d*
Successful retrieval attempts					
Massed practice	1199	52	145	1167	
Short-spaced practice	2198	52	535	2113	2.14
Long-spaced practice	2649	52	898	2375	1.70
Unsuccessful retrieval attempts					
Massed practice	2354	52	814	2232	
Short-spaced practice	3658	52	1412	3447	1.37
Long-spaced practice	3675	52	1512	3224	1.16

Here, overall, we see the same pattern of differences among the three ISI conditions. The table further shows more overall effort in the unsuccessful than in the successful retrieval attempts.

Figure 9 presents a line graph of the study-phase response latencies across the six repetitions in the three ISI conditions. Median values are presented instead of means due to a significant positive skew in the raw response latency data. This figure shows overall response times, regardless of the correctness of the response.

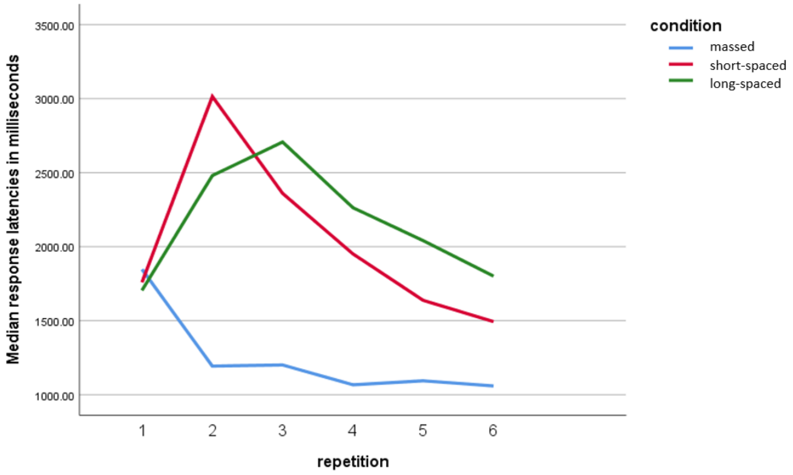


Figure 9: Median study-phase response latencies across the six repetitions

Overall, response latencies show a decrease across the repeated encounters; however, response times in the massed condition decrease quite dramatically in the early retrieval attempts, after which they do not decrease much because of a kind of a floor effect. In fact, the first true retrieval attempt, which occurs at repetition two, already exhibits a very low effort value in this condition. Responses in the two spaced conditions show much longer latencies across the repetitions, with the long-spaced condition continuing to elicit longer response latencies than those in the short-spaced condition until the very last repetition.

Figure 10 presents these response latencies separately for the long and short study duration conditions.

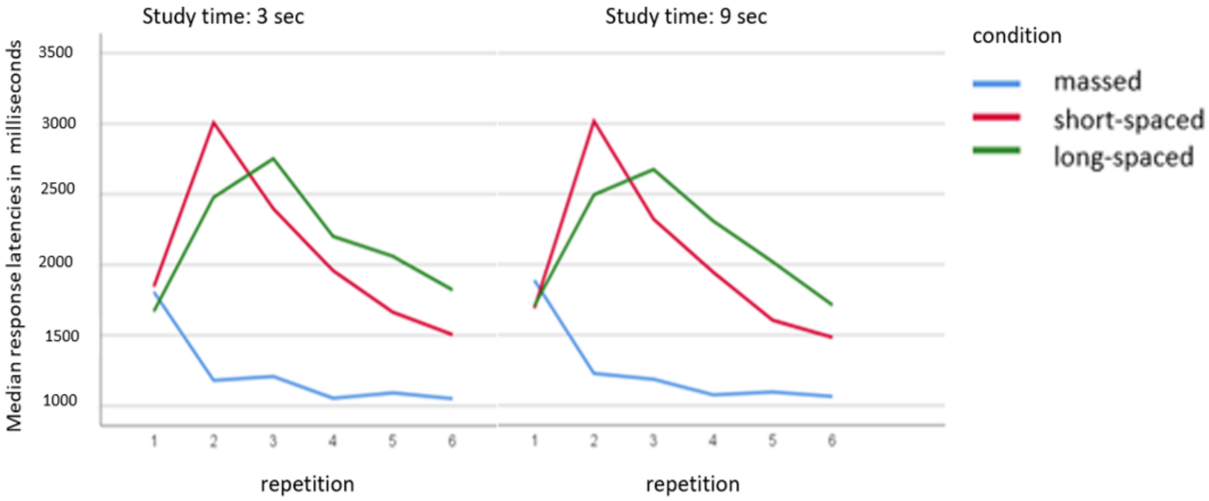
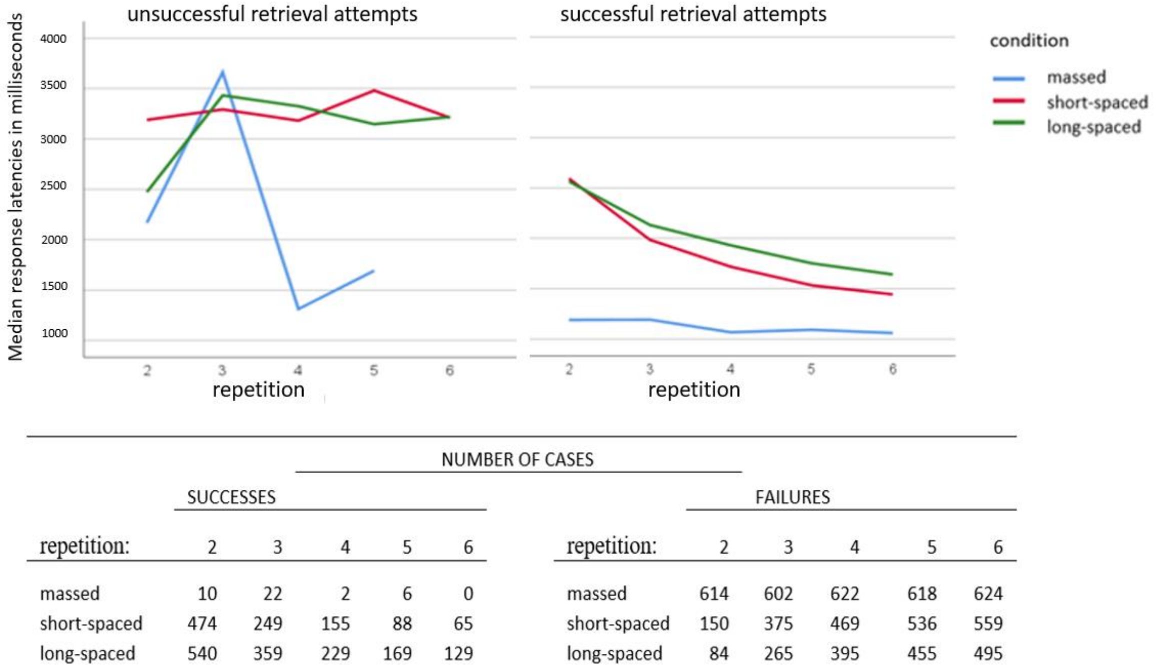


Figure 10: Response latencies in the short and long study duration conditions

The two study duration conditions appear to have produced very similar response latencies in the three conditions across the six repetitions. *Figure 11* presents response latencies separately for successful and unsuccessful retrieval attempts. It also presents the number of cases at each repetition in each ISI condition for successful and unsuccessful retrieval attempts. These numbers must be kept in mind when interpreting the trends in *Figure 11*. In this figure, the first encounter is excluded because it cannot be included in one of the graphs (the success graph, as all retrieval attempts for repetition one were, as expected, unsuccessful). This was done for ease of comparison across the two graphs. Further, two separate graphs are presented instead of a single graph because there is considerable overlap in the lines between the two study time conditions. Thus, *Figure 11* shows latencies only for true retrieval attempts, where the participants had studied each word before and therefore retrieval was possible, excluding retrieval attempts where the relevant translation had not yet been seen.



NUMBER OF CASES

	SUCCESSES						FAILURES				
repetition:	2	3	4	5	6	repetition:	2	3	4	5	6
massed	10	22	2	6	0	massed	614	602	622	618	624
short-spaced	474	249	155	88	65	short-spaced	150	375	469	536	559
long-spaced	540	359	229	169	129	long-spaced	84	265	395	455	495

Figure 11: Study-phase latencies in successful and unsuccessful retrieval attempts

The figure shows overall longer latencies for the unsuccessful retrieval attempts than for successful retrieval attempts. Further, while the latencies decrease quite steadily across repetitions in the successful retrieval attempts, though in a bit of a quadratic trend, latencies for the unsuccessful retrieval attempts do not seem to decrease across repetitions except in the massed condition, where the total number of unsuccessful retrieval attempts is very small. Repetition two seems to have produced similar effort between the two spaced conditions in the successful retrieval attempts at around two and a half seconds, however, the number of successful retrieval attempts in the long-spaced condition here is relatively small. The successful retrieval latencies further show a considerable difference between the massed condition and the two spaced conditions, the latter conditions producing considerably longer response times.

Study-phase response latencies: Inferential statistics. Growth curve modelling was used for initial exploration of changes in effort in the three conditions across repetitions as well as of how this may be affected by the time participants are allowed to study each word with its translation at each repetition. For this analysis, only experimental items were used because the control items did not repeat and thus cannot have a growth process. Further, latencies for the first encounters were removed from this analysis as well, as these do not represent true retrieval attempts (here, a search of the memory is performed in vain, as the translation for a given Finnish word has not been seen yet). *Figure 12* shows the growth trajectories across the three ISI conditions with the use of median values due to the positive skew in the distribution of latencies. Note that these latencies contain both correct and incorrect responses.

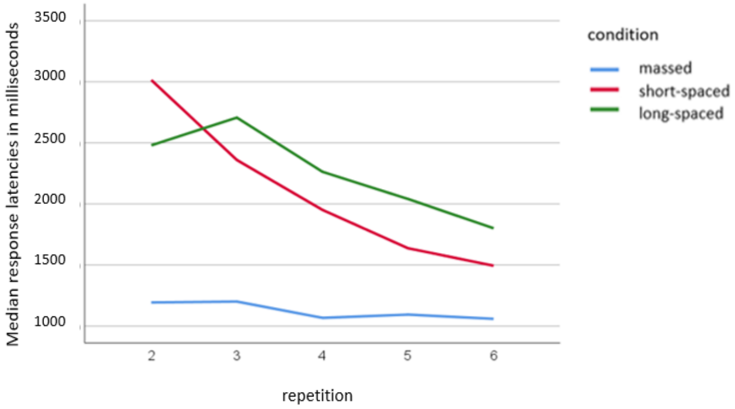


Figure 12: Growth in the latencies in the three conditions across repetitions

The distribution of residuals in response latencies was positively skewed (skewness = 5.265, SE_{skewness} = .025) and leptokurtic (kurtosis = 49.488, SE_{kurtosis} = .051). A natural log transformation was used to bring residuals to approximate more closely a normal distribution. Further, outliers above 3SD were removed (138 cases: 5 from the massed condition, 50 from the short-spaced condition, and 81 from the long-spaced condition). There were no outliers below 3SD. The resulting distribution was bell-shaped and followed the diagonal of the Q-Q plot quite closely (skewness = .562, SE_{skewness} = .026; kurtosis = .795, SE_{kurtosis} = .051).

A linear mixed-effects growth curve modeling analysis was used to explore change processes in retrieval effort across the five repeated retrieval attempts as well as how the trend may differ depending on retrieval success, study time duration, and ISI. However, because an initial analysis showed that there was a significant interaction between retrieval success and a linear and quadratic growth trajectories, $\chi_{(1)}$ = 65.839, p < .001 and $\chi_{(1)}$ = 4.767, p = .029, respectively), the latencies for the successful and unsuccessful retrievals were investigated separately. This also makes theoretical sense. In the field of psychology, only correct responses are usually investigated (e.g., Maddox et al., 2018), though this may be due to the fact that no feedback is usually provided in such studies and, consequently, effortful search of one’s memory that is unsuccessful still results in probable forgetting of the item in question in the absence of feedback. This may be different for effortful unsuccessful retrieval attempts that are followed by feedback (Kornell et al., 2009). However, while an investigation of latencies may be important even in the unsuccessful retrieval attempts in the present case, any growth processes will only be investigated in the successful retrieval attempts. The main reason for this is that in addition to capturing latencies in cases where a participant thought long and hard and still failed to produce the correct translation, the latencies in incorrect responses in the present experiment also capture situations where a participant did not recognize a word as repeated at all or did not attempt retrieval due to a quick estimation of how low the likelihood of success was. Here, some of the latencies at longer ISIs may actually often be shorter due to this and not to any effort processes while other latencies may be longer due to the ISI and its effect on effort processes, the two effects pulling in opposite directions. Thus, for instance, the upward growth in the long-spaced condition between repetitions two and three (see *Figure 12*) likely indicates that while upon the second repetition, which was separated from the initial encounter by a considerable amount of time and number of other items, many participants may not have recognized a given word as one they had studied before (or, if they did recognize it, did not attempt retrieval of its translation), upon repetition three, they may have been more likely to recognize the word and take the time to try to remember its translation, unsuccessful as this attempt may have been. Therefore, the amount of effort here depends on whether a retrieval attempt was undertaken at all as well as the actual effort of a search for the translation in one's memory. As it is impossible to disentangle these effects, analyzing latencies as a growth process may not be useful here. Additionally, one of the ISI conditions has too few observations to be useful in this analysis.

A multi-level framework was adopted to adjust for the nested structure of the data, as multiple data points were contributed by each of the participants. The intraclass correlation coefficient (ICC) for the effect of participant was .065, which indicates that roughly 6.5% of the variability in reading times can be attributed to the differences among the participants (Hayes, 2006). While this is a relatively small ICC, including the second level may still be safer than ignoring any, however small, dependency in the data, particularly since the software used allows such an analysis (Hayes, 2006). The fact that multiple encounters with the same Finnish words occurred both within and between participants further makes the words a potential level two variable within which encounters are nested. Finnish words produced an ICC of the same magnitude as participants (ICC = .065). The inclusion of both participants and words as random intercepts significantly improved model fit, $\chi_{(1)} = 2171.373$, $p < .001$. Both random intercepts were included. Model fit improvement was used as a measure of significance with Full Maximum Likelihood Estimation.

The inclusion of repetition as an independent variable significantly improved model fit, $\chi_{(1)} = 787.039$, $p < .001$. The addition of a quadratic term further significantly improved model fit, $\chi_{(1)} = 4.439$, $p = .035$, suggesting that the growth trajectory may not be linear. However, there was further a significant interaction between the quadratic term and condition, $\chi_{(1)} = 473.149$, $p < .001$, suggesting that the shape of the trajectory may differ depending on condition. Duration did not add a significant effect, $\chi_{(1)} = .098$, $p = .754$; there were further no other significant interactions, all *p*s > .05.

Because there was a significant condition by trend interaction, growth in the three conditions was examined separately. In each condition, there was a significant quadratic trend (massed: $\chi_{(1)} = 10.231$, $p = .001$; short-spaced: $\chi_{(1)} = 66.809$, $p < .001$; long-spaced: $\chi_{(1)} = 26.698$, $p < .001$). However, the conditions differed in terms of a cubic trend: while the massed condition exhibited a cubic trend ($\chi_{(1)} = 9.844$, $p = .002$), the other two did not (all *p*s > .05). However, the cubic trend in the massed condition might be an artefact of distractors being presented always after the second and, often, the third repetition, which may have produced slight amounts of forgetting between the respective repetitions in the massed condition, therefore, this trend may not be a reliable indication of changes in effort with repetition per se. Let us now turn to an investigation of the difference, at each repetition, in the amount of effort that a successful retrieval required. As can be seen in *Figure 12*, despite the quadratic trends, within the five true retrieval attempts, the effort in the long-spaced condition never decreased to the point of being equal to that in the short-spaced condition, which, in turn, never decreased to the point of being equal to the massed condition. This suggests that retrieval continued to be more effortful in the longer spaced condition than in the shorter spaced condition, even in later repetitions. *Figure 13* presents a growth curve that contains only those words for which successful retrieval attempts occurred at repetition two. Because it was nearly always the case that once a translation was correctly retrieved it continued to be correctly retrieved across later repetitions, *Figure 13* is a more pure illustration of how retrieval effort changed across the five repeated successful retrieval events in the three conditions.

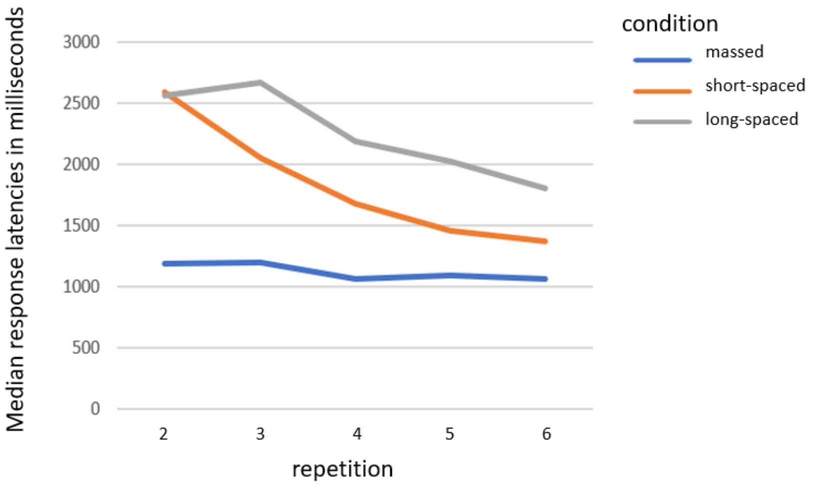


Figure 13: Response latencies across five successful retrieval attempts

No statistical analysis will be performed on the differences in these trajectories because of the considerable differences in the number of cases across the ISI conditions. Core statistical analyses will, instead, focus on the amount of effort, collapsed across repetitions, induced by the different levels of ISI as well as how this may interact with retrieval success rate and exposure duration, a variable that did not seem to affect the change across repetitions in the growth curve analysis latencies.

The sum of latencies across the five true retrieval attempts were used as the outcome variable to investigate any effects of ISI and presentation duration on response latencies during the study phase. The distribution of the residuals in the dependent variable was not normally distributed, (skewness = 2.183, SEskewness = .057, kurtosis = 9.998, SEkurtosis = .113). The natural log transformation was used to bring the distribution closer to a normal distribution. The resulting distribution of residuals was more nearly normal (skewness = .458, SEskewness = .057, kurtosis = .868, SEkurtosis = .113). The ICC for the effect of participant was .099, suggesting that roughly 10% of the variability in the dependent variable can be attributed to the differences between participants (Hayes, 2006). The ICC for the effect of the target words was .013, suggesting that roughly 1% of the variability in the dependent variable can be attributed to the target words. The inclusion of words as a random effect did not improve model fit and interfered with convergence, therefore this random effect was not included. The addition of participants as a random intercept significantly improved model fit, $\chi_{(1)} = 109.681$, $p < .001$. The addition of the number of correct retrieval attempts as a covariate significantly improved model fit, $\chi_{(1)} = 1124.907$, $p < .001$, indicating that there is, in fact, statistically significant difference in latencies between the successful and unsuccessful retrieval attempts. The addition of condition significantly improved model fit, $\chi_{(1)} = 697.172$, $p < .001$. However, there was also a significant interaction between condition and the number of correct retrieval attempts, $\chi_{(1)} = 19.141$, $p < .001$. *Figure 14* presents this interaction graphically.

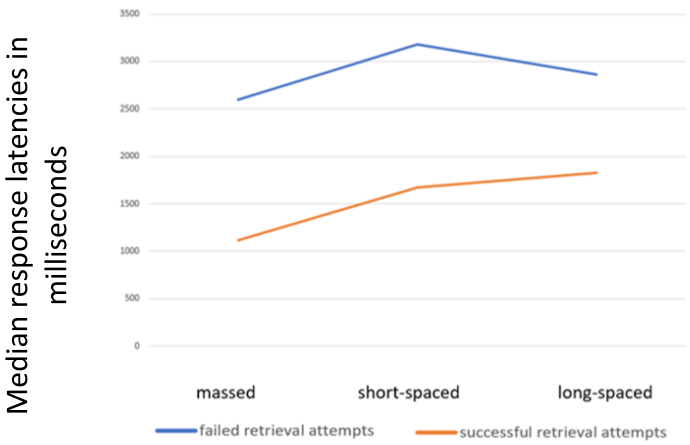
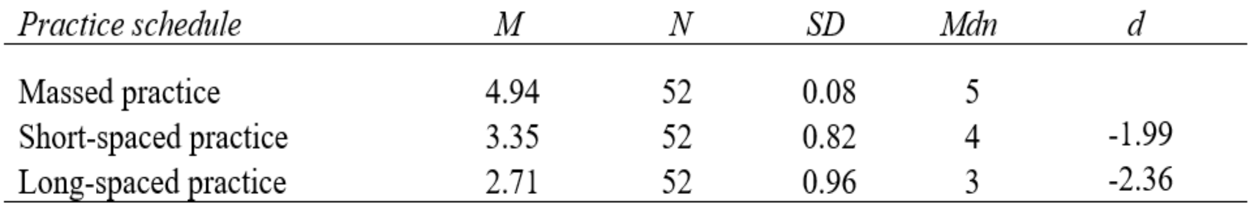


Figure 14: Response latencies as a function of condition and success of retrieval

Here, we see that failed retrieval attempts produced longer latencies overall and that spacing produced longer latencies as well. However, while effort appeared to grow monotonically across the levels of spacing in successful retrieval attempts, in unsuccessful retrieval attempts, the short-spaced condition appears to have produced slightly longer response latencies than the long-spaced condition. This might be due to the fact that more words were recognized as repeated in the short-spaced condition than in the long-spaced condition, in which case, more retrieval attempts (though unsuccessful in this case) were undertaken in the short-spaced condition, which shows up as more effort overall. Restricted Maximum Likelihood estimation was used to investigate these interactions due to the complexity of the model. Separate analyses were done for successful and unsuccessful retrieval attempts. The analyses showed a significant effect of ISI in both success conditions, all *p*s < .001. Parameter estimates with the long-spaced condition as the intercept were examined for more detailed information on how the three ISI conditions differed among themselves. This analysis showed that, in the successful responses, the massed and short-spaced conditions both significantly differed from the long-spaced condition (massed: $t_{(6761)}$ = -54.127, p < .001; short-spaced: $t_{(6749)}$ = -9.406, p < .001), the negative t values suggesting that both the massed and the short-spaced conditions received less effort than the long-spaced condition. However, in the unsuccessful attempts, the latencies in the short-spaced condition were significantly longer than those in the long-spaced condition ($t_{(2378)}$ = 4.722, p < .001). Further, the massed condition was not significantly different from the long-spaced condition ($t_{(2377)}$ = -1.673, p = .094). Thus, in unsuccessful attempts, retrieval effort was greatest in the shorter spaced condition while in the long-spaced condition, retrieval effort was almost of the same magnitude as that in the massed condition. This pattern may be explained in the same terms as the pattern in the graph: when a participant quickly estimated that they would not be able to produce a translation for a word they had not seen for a long time – or when they did not even recognize it as one they had studied before – they often gave a very quick “I don’t know” response. Thus, because in the short-spaced condition, the previous encounter was always a shorter time ago, here more retrieval attempts were undertaken (which means some effort was put into them) even if they were ultimately unsuccessful. A further analysis revealed that in the successful attempts, the massed condition produced significantly less effort than the short-spaced condition ($t_{(6000)}$ = -53.086, p < .001). Thus, the analysis of latencies has revealed a significant effect of lag on latencies in successful retrieval attempts, whereby retrieval effort increased with longer ISIs.

Study-phase retrieval success: Descriptive statistics. Table 23 presents the mean and median numbers of correct retrieval events during study phase in the three experimental conditions. The effect sizes here are calculated relative to the massed condition. Recall that there were a total of six repetitions per word and that upon the first repetition the word had not been presented before, therefore, "I don’t know” was the correct response. Thus, there were a total of five correct retrieval events possible out of the six total repetitions.

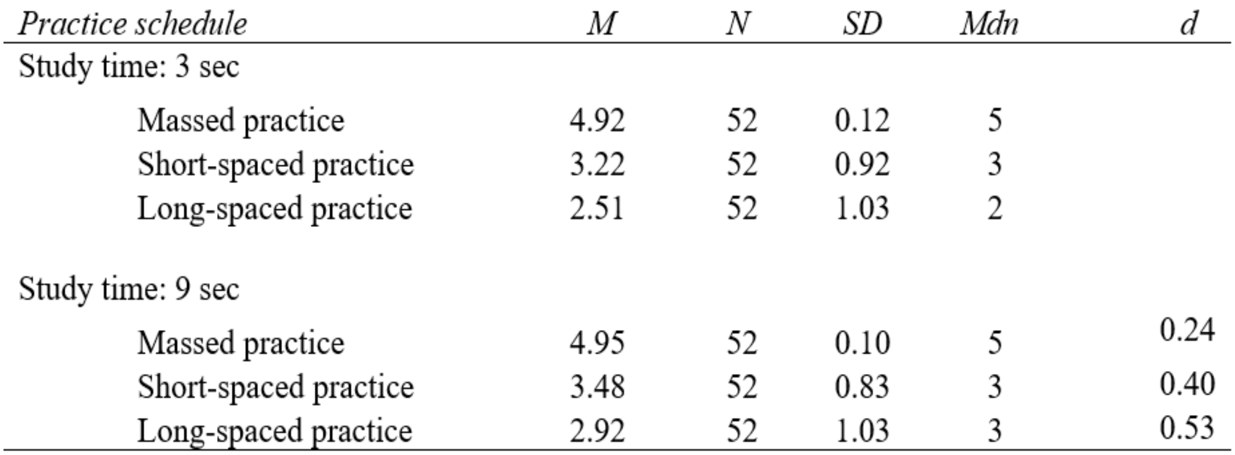
Table 23: *Correct retrieval events per experimental condition*



Practice schedule	*M*	*N*	*SD*	*Mdn*	*d*
Massed practice	4.94	52	0.08	5	
Short-spaced practice	3.35	52	0.82	4	-1.99
Long-spaced practice	2.71	52	0.96	3	-2.36

The retrieval attempts in the massed condition were almost always successful. The average number of successful retrieval attempts decreased with spacing such that in the short-spaced condition there were fewer successful retrieval events and in the long-spaced condition these were even fewer. The median values show a linear decrease in retrieval success across the spacing intervals while the means exhibit a bit of a quadratic trend, where the difference between the massed and short-spaced conditions is larger than that between the short- and long-spaced conditions. Table 24 presents these statistics separately in the two study-duration conditions (3 seconds vs. 9 seconds of studying a Finnish word with its translation). The effect sizes here represent differences between the short and long study time conditions.

Table 24: *Study-phase retrieval success in the short and long study time conditions*



Practice schedule	*M*	*N*	*SD*	*Mdn*	*d*
Study time: 3 sec					
Massed practice	4.92	52	0.12	5	
Short-spaced practice	3.22	52	0.92	3	
Long-spaced practice	2.51	52	1.03	2	
Study time: 9 sec					
Massed practice	4.95	52	0.10	5	0.24
Short-spaced practice	3.48	52	0.83	3	0.40
Long-spaced practice	2.92	52	1.03	3	0.53

The number of retrieval successes show a small benefit of longer study time, with this difference becoming larger the longer the spacing between repetitions.

Figure 15 presents a line graph that shows the growth in retrieval success across the repetitions in the three ISI conditions.

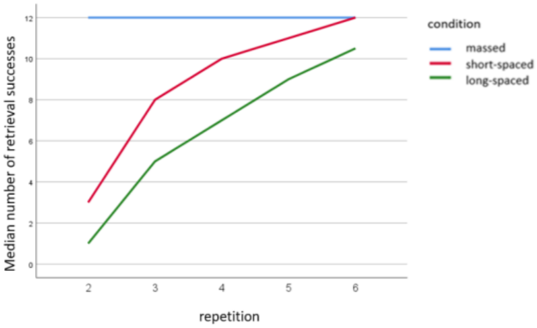


Figure 15: Successful retrievals at each repetition in the three conditions

The graph in *Figure 15* shows positive growth in the median number of successful retrieval attempts across the repetitions in the two spaced conditions; however, the conditions differ in the rate of such positive growth. In the massed condition, on the other hand, the median success value is at 100% from the very first retrieval attempt. In the short-spaced condition the median success value reached 100% only upon the last repetition and in the long-spaced condition the median success value never reached 100% within the five retrieval attempts. Further, the growth in success in the two spaced conditions looks to be almost parallel, with the short-spaced condition exhibiting a higher rate of success across the repetitions. *Figure 16* presents the growth in retrieval success separately in the two study time conditions. Two graphs are presented side by side due to a considerable overlap in the lines.

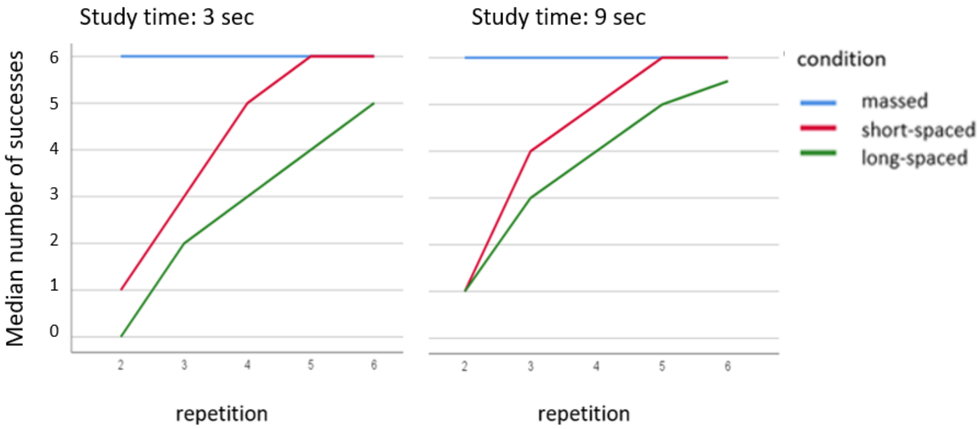


Figure 16: Growth in retrieval successes in the two study time conditions

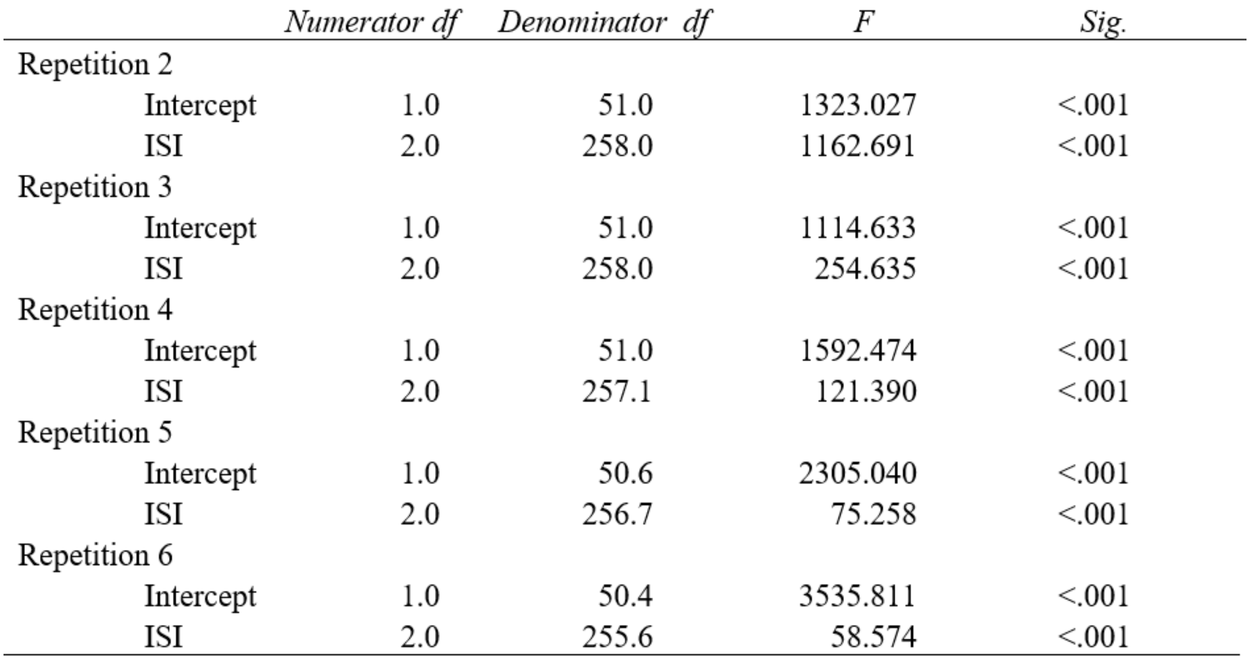
This graph shows that longer study time was beneficial for both spaced conditions; however, it looks to be a bit more beneficial for the long-spaced condition.

Study-phase retrieval success: Inferential statistics. To investigate how retrieval success changes with repetition in the three ISI conditions and whether this is affected by presentation duration, the number of successful retrieval attempts across repetitions was used as the dependent variable in a growth curve analysis. While the number of successes was a count variable consisting of 5 possible values (the lowest acceptable number for doing linear analyses on count data), residuals presented almost a normal distribution (skewness = -.162, SE_{skewness} = .062; kurtosis = .256, SE_{kurtosis} = .124) with the exception of 4 outliers in the lower values that were above 3SD. These outliers were removed and the distribution became even more nearly normal (skewness = -.065, SE_{skewness} = .062; kurtosis = -.027, SE_{kurtosis} = .124). The Q-Q plot further showed that the data closely followed the diagonal. Further, the histogram was bell-shaped as well, suggesting that a linear analysis was an acceptable option. A linear mixed effects growth curve model was fitted. The ICC for the effect of participant was .087, suggesting that roughly 9% of the variance in the dependent variable was due to the effect of participant differences. The addition of participants as random intercepts significantly improved model fit, $\chi_{(1)} = 68.252$, $p < .001$. For these reasons, participants were included as random effects in the analysis. Full Maximum Likelihood Estimation was used to investigate these growth processes.

The inclusion of repetition as an independent variable significantly improved model fit, $\chi_{(1)} = 431.771$, $p < .001$. The trend was further significantly quadratic, $\chi_{(1)} = 50.808$, $p < .001$. The addition of a cubic term negatively affected model fit. Therefore, only the quadratic term was retained. The addition of condition as an independent variable significantly improved model fit, $\chi_{(1)} = 1035.410$, $p < .001$. There was further a significant interaction between condition and repetition $\chi_{(1)} = 551.684$, $p < .001$, as well as between condition and the quadratic trend $\chi_{(1)} = 85.926$, $p < .001$, suggesting that in addition to a difference in slopes, or the rate of growth, the conditions also differed in the shape of these trajectories. For this reason, the growth trajectories as well as any effects of study on these trajectories were examined separately in the three ISI conditions. The alpha level was adjusted accordingly for all subsequent analyses: $\alpha = .05/8 = .006$. In the massed condition, there was a significant positive slope for the effect of repetition, $\chi_{(1)} = 13.571$, $p < .001$, however, the addition of a quadratic term did not improve the model, $\chi_{(1)} = 1.094$, $p > .05$. There was, further, no effect of presentation duration, $\chi_{(1)} = 2.135$, $p = .144$. There was further no significant interaction between study time and the linear and quadratic trends, all *p*s > .05. In the short-spaced condition, there was also a significant positive slope, $\chi_{(1)} = 482.042$, $p < .001$, and there was a significant quadratic trend, $\chi_{(1)} = 128.631$, $p < .001$ but no cubic trend, $\chi_{(1)} = 2.415$, $p > .05$. There was, further, a significant positive effect of presentation duration, $\chi_{(1)} = 13.521$, $p < .001$ but no significant interaction between presentation duration and the linear trend, $\chi_{(1)} = 4.637$, $p = .031$, nor the quadratic trend, $\chi_{(1)} = .429$, $p > .05$. An examination of the parameter estimates suggests that, on average, longer study time produced learning of .78 more words than shorter study time in this condition. Recall that raw numbers of words are used as the dependent variable in this analysis, therefore, the slope can be easily interpreted as the mean difference in the number of words successfully translated. In the long-spaced condition, the addition of repetition significantly improved model fit, $\chi_{(1)} = 503.060$, $p < .001$, and there was a significant quadratic trend, $\chi_{(1)} = 83.011$, $p < .001$ but no cubic trend, $\chi_{(1)} = 1.228$, $p > .05$. There was, further, a significant positive effect of study time, $\chi_{(1)} = 33.987$, $p < .001$, but no significant interaction between study time and the linear trend, $\chi_{(1)} = .223$, $p > .05$, or the quadratic trend, $\chi_{(1)} = .982$, $p > .05$. An examination of the parameter estimates suggests that, on average, longer study time produced learning of .48 more words than the shorter study time in this ISI condition.

The analyses presented above suggest that there was overall positive growth in the number of successful retrieval attempts across the repetitions and also that the steepness of this positive slope depended on the ISI condition. To investigate the effect of ISI condition on retrieval success, each repetition was investigated separately. Five omnibus analyses were run – one for each repetition – and the parameter estimates were investigated. Linear mixed effects modeling was used with REML due to the complexity of the model. The short-spaced condition was the reference category against which the effects of the other two conditions were tested for significance. Table 25 presents the results of the omnibus tests across the repetitions.

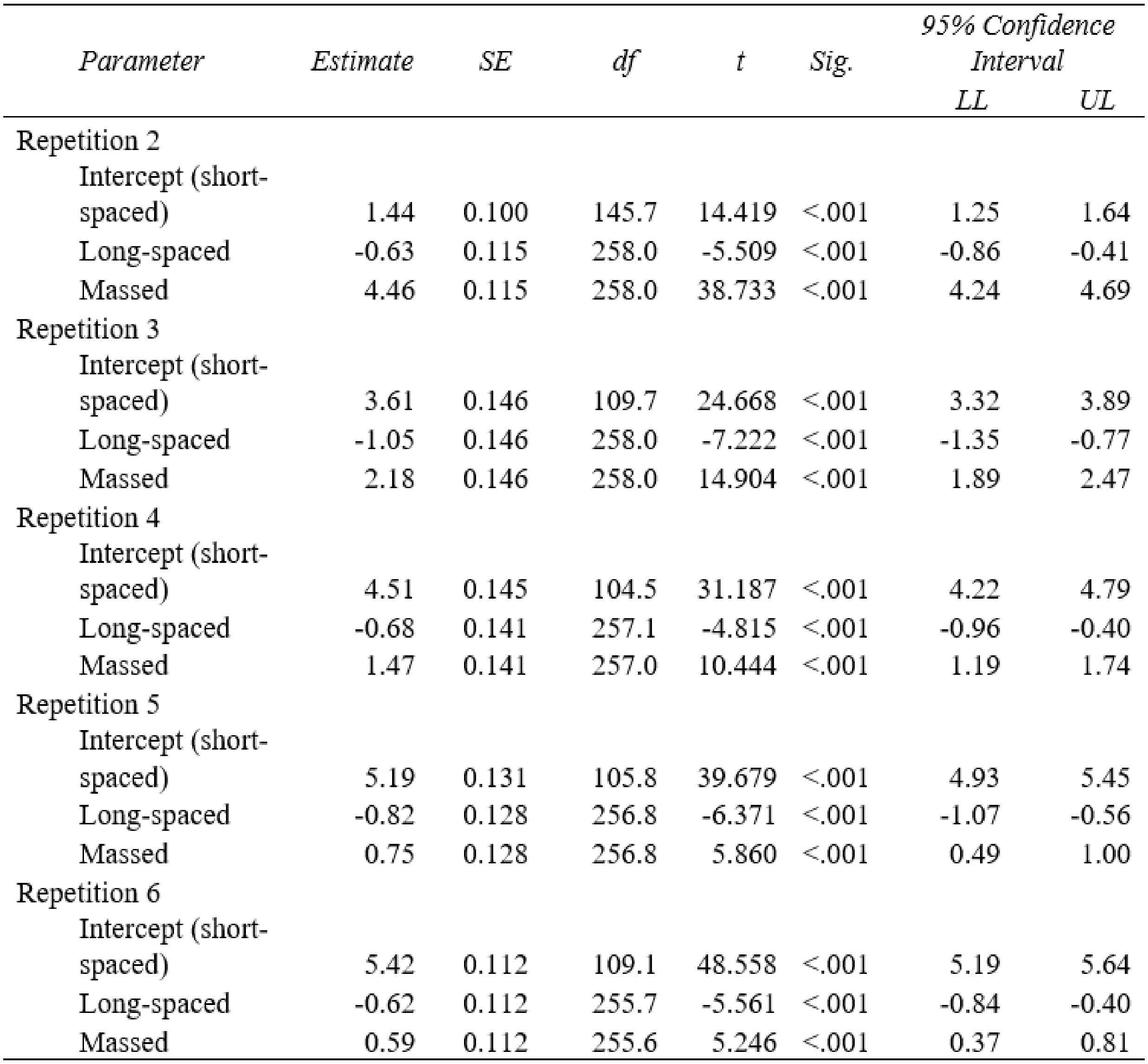
Table 25: *The effect of ISI on retrieval success at the five repetitions*



	Numerator df	*Denominator df*	*F*	*Sig.*
Repetition 2				
Intercept	1.0	51.0	1323.027	<.001
ISI	2.0	258.0	1162.691	<.001
Repetition 3				
Intercept	1.0	51.0	1114.633	<.001
ISI	2.0	258.0	254.635	<.001
Repetition 4				
Intercept	1.0	51.0	1592.474	<.001
ISI	2.0	257.1	121.390	<.001
Repetition 5				
Intercept	1.0	50.6	2305.040	<.001
ISI	2.0	256.7	75.258	<.001
Repetition 6				
Intercept	1.0	50.4	3535.811	<.001
ISI	2.0	255.6	58.574	<.001

Here we see that there was a significant difference between at least two of the groups (the alternative hypothesis for the omnibus test) in each repetition, all *p*s < .001. Table 26 presents the parameter estimates that provide information about differences among the three conditions. Here, the short-spaced condition is used as the reference category and, therefore, all comparisons are made against this condition. Recall that the data represent raw counts of words that were correctly retrieved during the study phase at each repetition in the three conditions. For this reason, the intercept can be interpreted in terms of the number of translations correctly retrieved in the short-spaced condition (the reference category) and each slope can be interpreted in terms of the difference, in raw numbers of words correctly retrieved, between the short-spaced condition and each of the other two conditions.

Table 26: *Parameter estimates for the effect of ISI on study-phase retrieval success*



Parameter	*Estimate*	*SE*	*df*	*t*	*Sig.*	*95% Confidence Interval*	
						LL	*UL*
Repetition 2							
Intercept (short-spaced)	1.44	0.100	145.7	14.419	<.001	1.25	1.64
Long-spaced	-0.63	0.115	258.0	-5.509	<.001	-0.86	-0.41
Massed	4.46	0.115	258.0	38.733	<.001	4.24	4.69
Repetition 3							
Intercept (short-spaced)	3.61	0.146	109.7	24.668	<.001	3.32	3.89
Long-spaced	-1.05	0.146	258.0	-7.222	<.001	-1.35	-0.77
Massed	2.18	0.146	258.0	14.904	<.001	1.89	2.47
Repetition 4							
Intercept (short-spaced)	4.51	0.145	104.5	31.187	<.001	4.22	4.79
Long-spaced	-0.68	0.141	257.1	-4.815	<.001	-0.96	-0.40
Massed	1.47	0.141	257.0	10.444	<.001	1.19	1.74
Repetition 5							
Intercept (short-spaced)	5.19	0.131	105.8	39.679	<.001	4.93	5.45
Long-spaced	-0.82	0.128	256.8	-6.371	<.001	-1.07	-0.56
Massed	0.75	0.128	256.8	5.860	<.001	0.49	1.00
Repetition 6							
Intercept (short-spaced)	5.42	0.112	109.1	48.558	<.001	5.19	5.64
Long-spaced	-0.62	0.112	255.7	-5.561	<.001	-0.84	-0.40
Massed	0.59	0.112	255.6	5.246	<.001	0.37	0.81

Table 26 shows that there were significantly more successes in the massed condition than in the short-spaced condition at each repetition. It also shows that there were significantly fewer successes in the long-spaced condition than in the short-spaced condition at each repetition. This pattern of results obviates the need for a separate comparison between the massed and the long-spaced conditions as these conditions have significant slopes in opposite directions from the intermediate short-spaced condition and, therefore, we can conclude that they, too, are significantly different from each other.

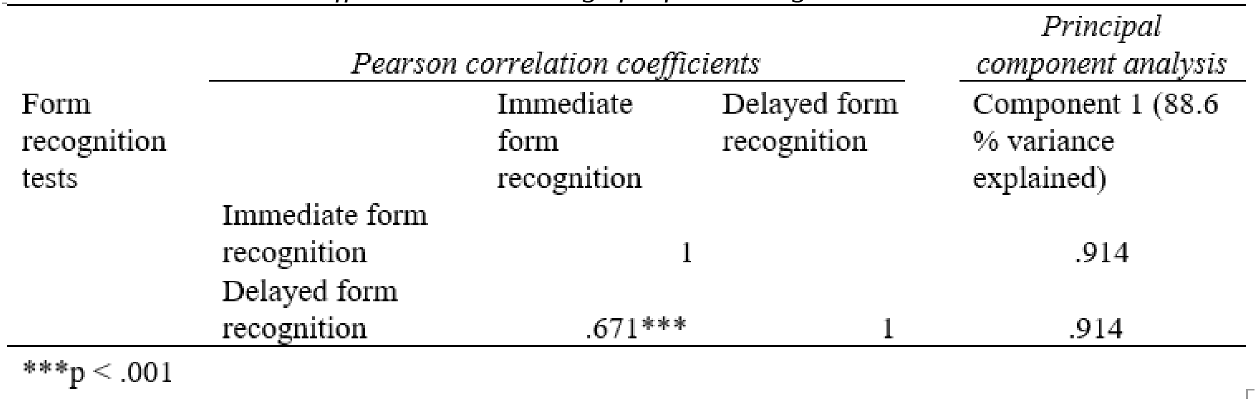
Thus, the analyses presented above have shown that the number of successes grew across the repetitions throughout the study phase, although the rate of this growth slowed in later repetitions (the quadratic trend), that growth in the three ISI conditions differed significantly in the number of successful retrievals, and that this difference did not disappear with repeated encounters throughout the study phase. The analyses further showed a significant positive effect of presentation duration on study-phase retrieval success in the two spaced conditions but not in the massed condition.

Moderated mediation analyses

The results of the previous analyses have shown that spacing repeated retrieval practice more widely results in superior learning outcomes. It further makes the study-phase retrieval process more effortful but also less successful. To answer the third research question that asks whether the dual mechanism of successful effortful retrieval underlies the effects of ISI on learning outcomes and whether study time moderates this relationship, two moderated mediation analyses were performed with the SPSS PROCESS 4.3 macro (Hayes, 2018).

Moderated parallel mediation analyses. Because learning outcomes were measured with multiple posttests, data reduction was performed to reduce the six tests to fewer dependent variables. Based on correlations, theoretical reasons, and principle component analyses, three dependent variables emerged. These combined together (1) the immediate and delayed form- recognition tests, (2) the two immediate meaning tests, and (3) the two delayed meaning tests. The three resulting tests will be named, respectively, the form-recognition tests, the immediate meaning tests, and the delayed meaning tests. Tables 27-29 present the bivariate two-tailed correlations between each member of a pair as well as loadings of each pair of tests on their corresponding extracted component.

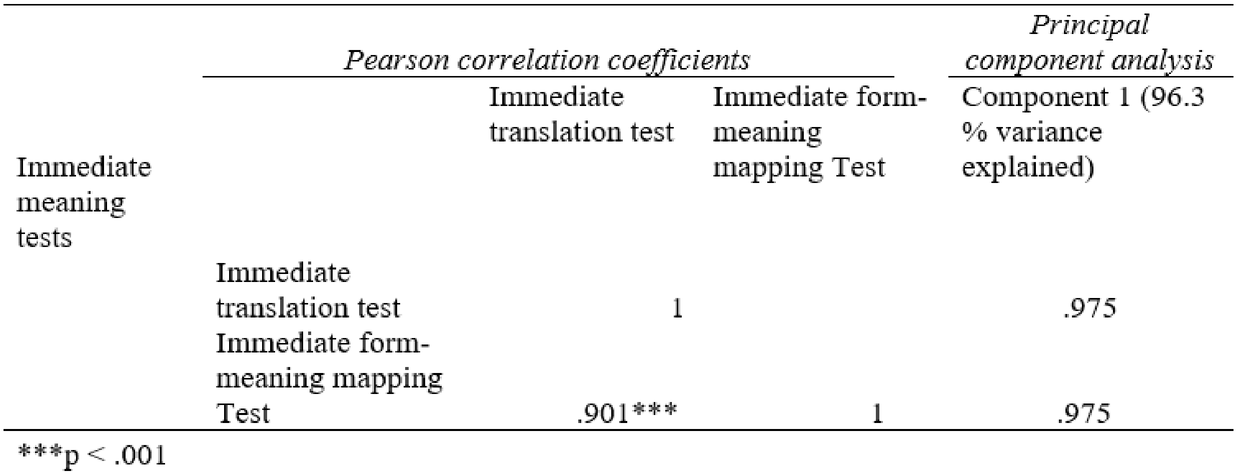
Table 27: *Correlation coefficients and loadings for form-recognition tests*



	Pearson correlation coefficients			*Principal component analysis*
Form recognition tests		Immediate form recognition	Delayed form recognition	Component 1 (88.6 % variance explained)
	Immediate form recognition	1		.914
	Delayed form recognition	.671***	1	.914

***p < .001

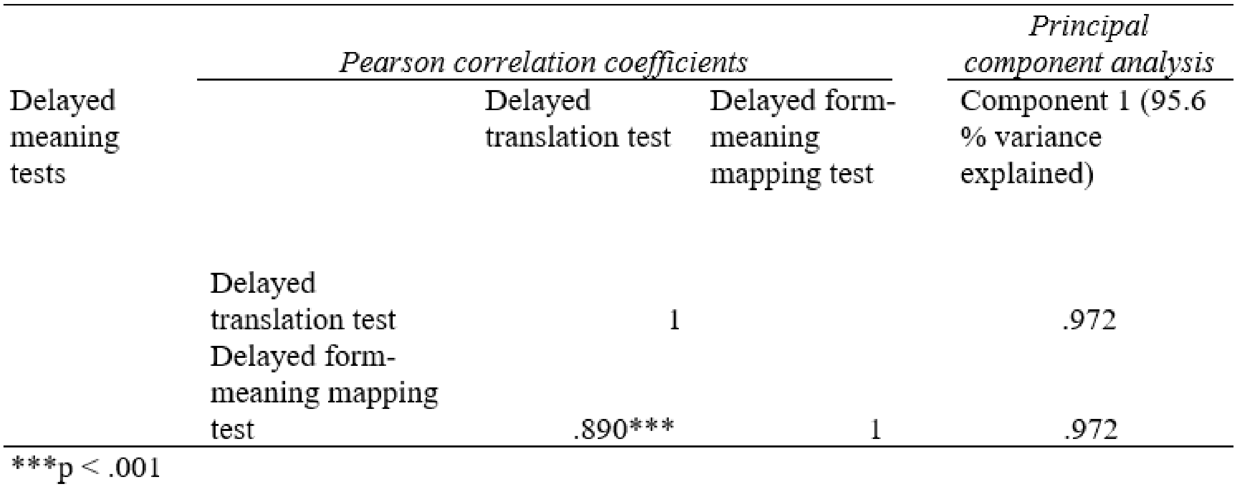
Table 28: *Correlation coefficients and loadings for immediate meaning tests*



	Pearson correlation coefficients			*Principal component analysis*
Immediate meaning tests		Immediate translation test	Immediate form-meaning mapping Test	Component 1 (96.3 % variance explained)
	Immediate translation test	1		.975
	Immediate form-meaning mapping Test	.901***	1	.975

***p < .001

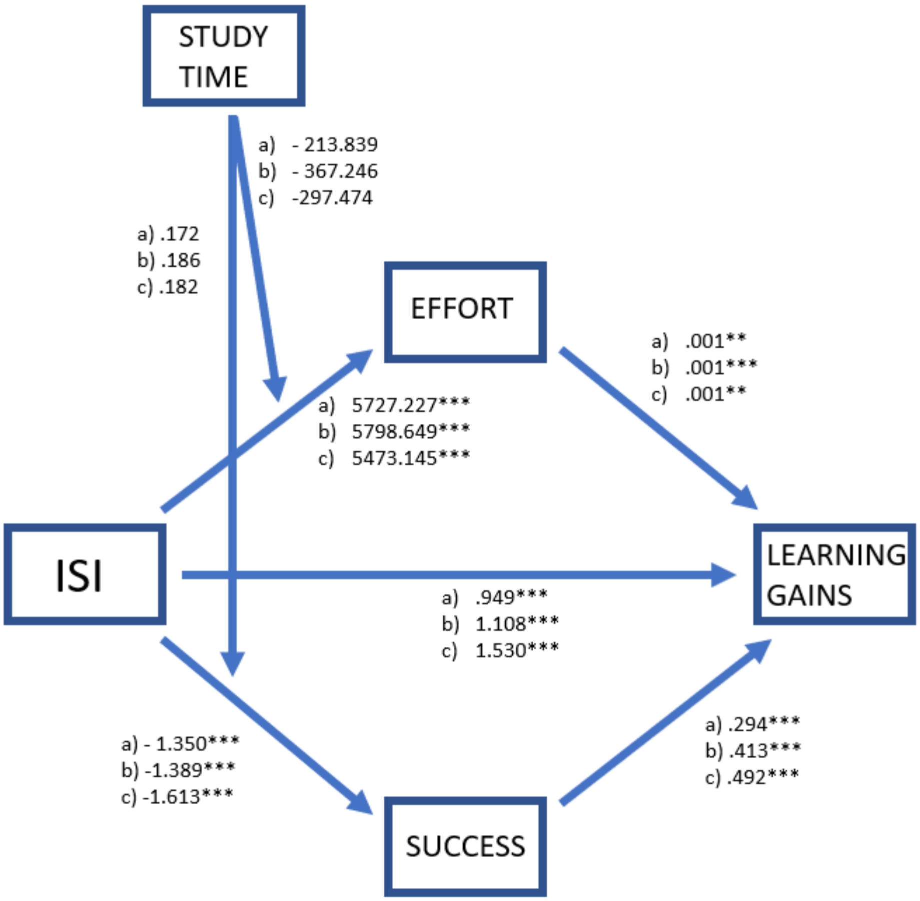
Table 29: *Correlation coefficients and loadings for delayed meaning tests*



	Pearson correlation coefficients			*Principal component analysis*
Delayed meaning tests		Delayed translation test	Delayed form-meaning mapping test	Component 1 (95.6 % variance explained)
	Delayed translation test	1		.972
	Delayed form-meaning mapping test	.890***	1	.972

***p < .001

Each table shows quite high loadings, suggesting that the corresponding test pair likely measures the same underlying construct. Because multiple models were run on the same or related data, the alpha level was corrected accordingly. Further, robust tests were run to ensure against any violations of normality. Thus, bootstrapped 99% confidence intervals were used with 10,000 bootstrap samples. An initial model investigated whether the two mechanisms of success and effort underlie any effects of lag in the present results as well as whether the operation of these two mechanisms as a function of ISI is affected by study time. The moderated parallel mediation included study-phase retrieval effort and success and the two mediators and study time as the moderator of the relationship between ISI and the two mediators (model 7). This model was tested with each of the three tests (each of which combined a pair of tests as discussed above). Further, time of delayed test was included as a covariate in the form-recognition test and the delayed meaning test because each of these two tests contained scores from a delayed test. *Figure 17* presents the conceptual structure of this analysis with obtained coefficients for each of the three tests.



**p < .01, ***p < .001

Figure 17: Conceptual structure for the moderated parallel mediation analysis

The form-recognition test. The coefficients for the form-recognition test show that, as found in earlier analyses, ISI had a significant positive effect on learning outcomes as well as a significant positive effect on effort and a significant negative effect on study-phase retrieval success. Additionally, the coefficients for the effect of study time show that this variable does not have a significant effect on the relationship between ISI and the two mediators of retrieval effort and success. This means that effort increases and success decreases across the three levels of ISI and these trends are not significantly affected by how much time a learner is given for study of a given Finnish word with its translation. The coefficients further show a significant positive effect of successful retrieval of a word’s meaning during study on form-recognition posttest scores and a significant positive effect of effort on these same scores. This means that both retrieval effort and retrieval success positively affect learning, which is in line with the predictions of the reminding account. Note that both effort and success are modeled here as main effects. However, the effect of one may depend on the level of the other, thus, the effect of effort on learning may depend on whether or not retrieval is successful, as proposed by the dual mechanism account under investigation. Whether this is the case will be explored in a subsequent analysis.

The tests of the indirect effects showed significant mediation by retrieval success as a negative effect across the two levels of study time: β = -.3463, bootstrapped standard error = .0820, 99% bootstrapped confidence interval [-.5754, -.1420] for short presentation duration; and β = -.2957, bootstrapped standard error = .0754, 99% bootstrapped confidence interval [-.5180, -.1224] for long presentation duration. This suggests that, despite the fact that there was no nonmonotonicity in the form-recognition scores as a function of lag in the present results, a negative effect of longer ISI was still present and operated through consequent lower study-phase retrieval success, which was true for both levels of presentation duration. Thus, lower levels of study-phase retrieval success significantly mediated negative effects on learning of wider spacing between repetitions, regardless of how long a given word was studied for.

The tests of the indirect effects further showed significant mediation by retrieval effort as a positive effect across the two levels of study time: β = .1689, bootstrapped standard error = .0528, 99% bootstrapped confidence interval [.0474, .3194] for short presentation duration; and β = 1624, bootstrapped standard error = .0484, 99% bootstrapped confidence interval

[.0472, .3018] for long presentation duration. This means that increased effort that resulted from spacing retrieval attempts more widely was beneficial for learning to recognize the target words. However, there was no significant overall moderated mediation process, Index of Moderated Mediation = -.0065, bootstrapped standard error = .0232, 99% bootstrapped confidence interval [-.0780, .0542], indicating that the operation of the two underlying mechanisms of retrieval effort and success did not depend on whether the Finnish words and their English translations were presented for 3 or 9 seconds after each retrieval attempt.

The immediate meaning test. The coefficients in *Figure 17* show a significant positive direct effect of ISI on the immediate meaning scores. The tests of the simple effects of ISI on effort and success as well as the moderating effects of duration on these variables are not affected by what outcome test is the dependent variable in any given model and will, therefore, be similar for the immediate meaning scores to those presented in the previous analysis of form-recognition scores as well as in the following analysis of the delayed meaning posttest scores. However, the entire model needs to be tested for each of the outcome tests because of the complexity of the underlying relationships. Therefore, despite being almost redundant, coefficients for the entire model are presented in *Figure 17*, for consistency, for each of the three outcome tests. These coefficients may look slightly different among the three outcome tests due to bootstrapping. However, the difference should be very small and should not affect interpretation. The effects of effort and success, however, will be different, as we have a different outcome variable. These coefficients show a small but significant positive effect of retrieval effort on the immediate meaning scores as well as a significant positive effect of successful retrieval on these scores. This means that for the immediate meaning scores, both retrieval effort and success have a significant positive effect, again in line with the predictions of the dual mechanism account.

The tests of the indirect effects on immediate meaning scores showed significant mediation by retrieval success as a negative effect across the two levels of study time: β = -.4972, bootstrapped standard error = .0794, 99% bootstrapped confidence interval [-.7238, -.3085] for short presentation duration and β = -.4204, bootstrapped standard error = .0713, 99% bootstrapped confidence interval [-.6220, -.2576] for long presentation duration. This suggests that, despite the fact that there was no nonmonotonicity in the immediate meaning scores as a function of lag in the present experiment, a negative effect of longer ISI was still present and operated through consequent lower study-phase retrieval success, which was true for both levels of presentation duration. Thus, lower levels of study-phase retrieval success significantly mediated negative effects on learning of wider spacing between repetitions, regardless of how long a given word was studied for.

The test of the indirect effects further showed significant mediation by retrieval effort as a positive effect across the two levels of study time: β = .1703, bootstrapped standard error = .0470, 99% bootstrapped confidence interval [.0568, .3038] for short presentation duration and β = 1588, bootstrapped standard error = .0414, 99% bootstrapped confidence interval [.0541, .2734] for long presentation duration. However, as with the form-recognition results, there was no significant overall moderated mediation process, Index of Moderated Mediation = -.0115, bootstrapped standard error = .0210, 99% bootstrapped confidence interval [-.0756, .0406], indicating that the operation of the two underlying mechanisms of retrieval effort and success did not depend on whether the Finnish words and their English translations were presented for 3 or 9 seconds after each retrieval attempt.

The delayed meaning test. The coefficients in *Figure 17* show a significant positive direct effect of ISI on the delayed meaning scores. There is also a significant positive effect of study-phase retrieval effort and success on these scores. The tests of the indirect effects on the delayed meaning scores showed significant mediation by retrieval success as a negative effect across the two levels of study time: β = -.7032, bootstrapped standard error = .1047, 99% bootstrapped confidence interval [-.9909, -.4530] for short presentation duration and β = -.6135, bootstrapped standard error = .0930, 99% bootstrapped confidence interval [-.8675, -.4021] for long presentation duration. Here, again, despite the fact that there was no nonmonotonicity in the delayed meaning scores as a function of lag in the present results – in fact, the delayed posttests showed a lag effect, whereby scores in the long-spaced condition were actually significantly higher than scores in the short-spaced condition – a negative effect of longer ISI was still present and operated through lower study-phase retrieval success, across the two levels of presentation duration. Thus, here again, lower levels of study-phase retrieval success significantly mediated negative effects on learning of wider spacing between repetitions, regardless of how long a given word was studied for.

The test of the indirect effects did not show significant mediation by retrieval effort and this was true across the two levels of study time: β = .1249, bootstrapped standard error = .0496, 99% bootstrapped confidence interval [.-0026, .2572] for short presentation duration and β = 1177, bootstrapped standard error = .0505, 99% bootstrapped confidence interval [-.0023, .2617] for long presentation duration. Further, as in the previous two tests, there was no significant overall moderated mediation process, Index of Moderated Mediation = .0896, bootstrapped standard error = .0508, 99% bootstrapped confidence interval [-.0293, .2311], indicating that the operation of the two mechanisms of retrieval effort and success was not affected by whether the Finnish words and their English translations were presented for 3 or 9 seconds after each retrieval attempt.

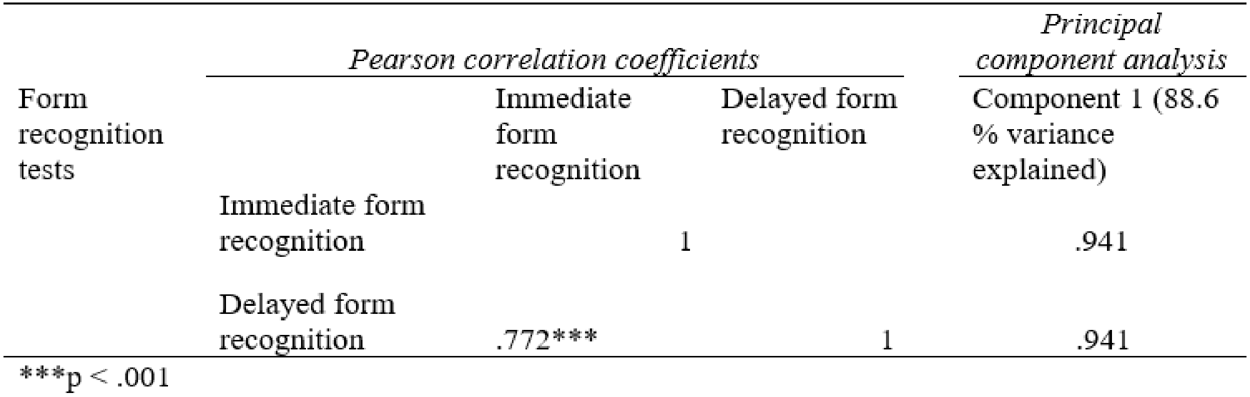
The moderated parallel mediation analyses showed no significant moderated mediation in any of the three sets of vocabulary scores, suggesting that study time did not affect the operation of the investigated underlying mechanisms of retrieval effort and success in the present study. In all three tests, retrieval success significantly mediated negative effects of ISI on learning outcomes. Thus, despite a failure to capture a nonmonotonic function of lag in learning outcomes in the present study, increasing the ISI did, in fact, have a negative effect on learning outcomes and this effect operated through a lower rate of study-phase retrieval success.

Study-phase retrieval effort did not have a significant main effect on learning, nor did it mediate the benefits of ISI, in the delayed meaning test scores, although it had both effects on the other two tests. On the surface, this latter finding is surprising and seems to suggest that higher amounts of effort are not beneficial for learning meanings of L2 words in the long term. However, because the proposed underlying mechanism is essentially an interaction between retrieval effort and success – that is, what underlies benefits of ISI is a mechanism of effortful successful retrieval – the main effect of effort may not be a stable effect and may, therefore, depend on the level of retrieval effort. The question whether the positive effects of retrieval effort are conditional on the level of retrieval success will be tested in the following moderated mediation analysis.

Mediation by retrieval effort moderated by retrieval success (a moderated mediation analysis). Retrieval effort was chosen as the mediator of the relationship between spacing and learning. Retrieval success was chosen as a moderator of this mediation. The reason for the choice of the mediator was theoretical. Because retrieval effort is known to promote word learning (Pyc & Rawson, 2009) and the amount of attentional engagement has been shown to mediate the benefits of spacing study of L2 vocabulary learning (Koval, 2019), it is an interesting question whether the benefits of increased effort that results from longer ISIs in retrieval practice are conditional on higher levels of retrieval success. It is further interesting to know whether this holds in the presence of feedback that follows each retrieval attempt. Provision of feedback after each retrieval attempt is a more usual situation for second language vocabulary learning. The moderated parallel mediation analysis showed that despite the fact that a nonmonotonic function was not observed in the learning outcomes, failure of study-phase retrieval that resulted from spacing retrieval attempts more widely still had a negative effect on learning. It is an important question whether retrieval success rate moderates beneficial effects of retrieval effort on learning and may thus constitute a limitation on how widely we may space retrieval practice even in the presence of feedback. Significant mediation in the present case would mean that spacing retrieval practice more widely positively affects learning outcomes because it increases retrieval effort, which, in turn, leads to better learning. Significant moderation of this mediation by retrieval success would mean that the benefits of effort (the mediator) may be conditional on retrieval success (the moderator) and, therefore, repetitions should not be spaced so widely that it negatively affects retrieval success, even in the presence of feedback.

Because study time was shown not to moderate the relationship between ISI and study-phase retrieval effort and success, participants' scores were collapsed across the levels of this variable for this analysis. Tables 30-32 present the bivariate two-tailed correlations between the members of each pair of tests as well as loadings of each pair of tests on their corresponding extracted component.

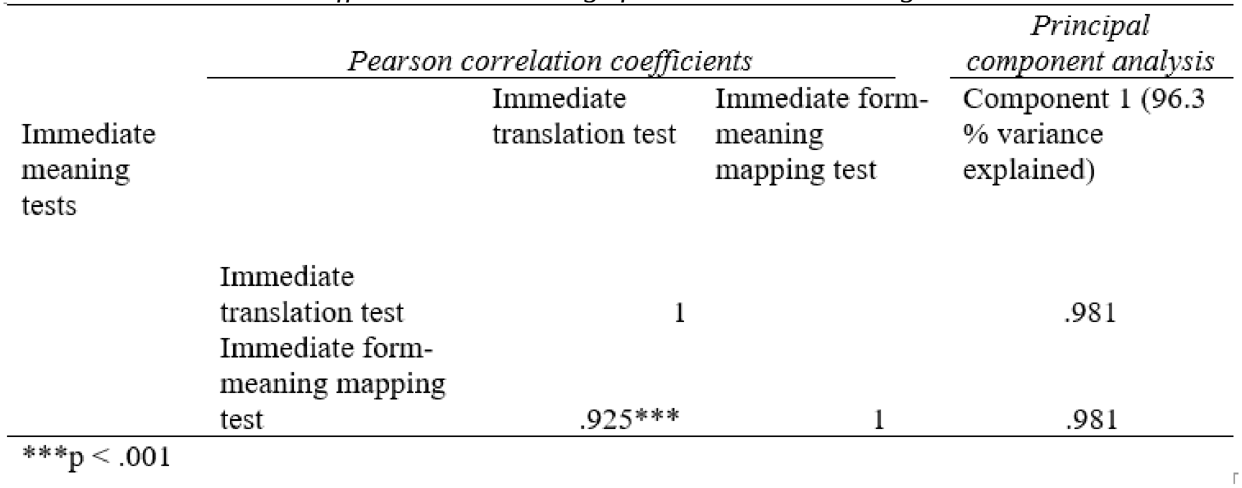
Table 30: *Correlation coefficients and loadings for form-recognition tests*



		Pearson correlation coefficients		*Principal component analysis*
Form recognition tests		Immediate form recognition	Delayed form recognition	Component 1 (88.6 % variance explained)
	Immediate form recognition	1		.941
	Delayed form recognition	.772***	1	.941

***p < .001

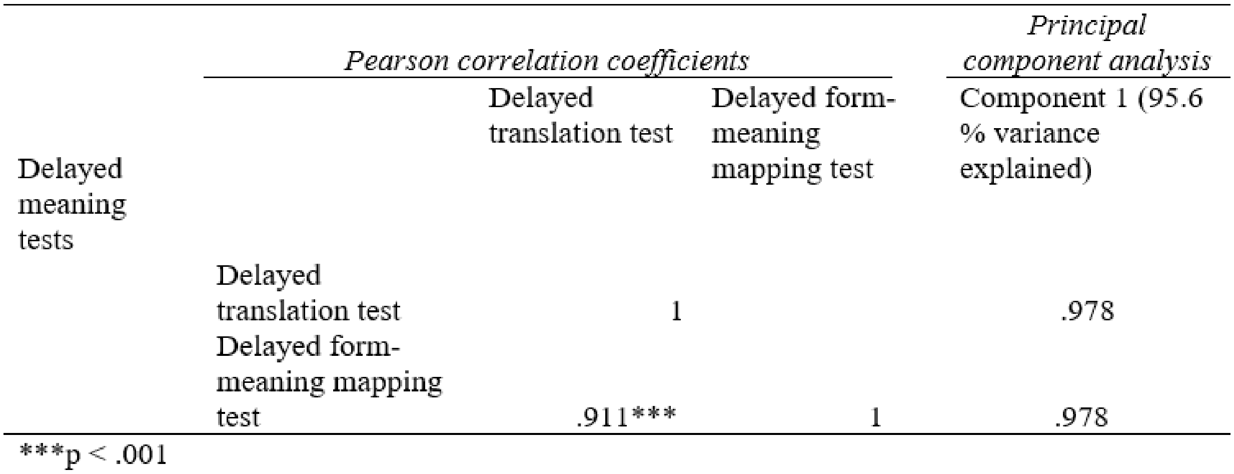
Table 31: *Correlation coefficients and loadings for immediate meaning tests*



		Pearson correlation coefficients		*Principal component analysis*
Immediate meaning tests		Immediate translation test	Immediate form-meaning mapping test	Component 1 (96.3 % variance explained)
	Immediate translation test	1		.981
	Immediate form-meaning mapping test	.925***	1	.981

***p < .001

Table 32: *Correlation coefficients and loadings for delayed meaning tests*



		Pearson correlation coefficients		*Principal component analysis*
Delayed meaning tests		Delayed translation test	Delayed form-meaning mapping test	Component 1 (95.6 % variance explained)
	Delayed translation test	1		.978
	Delayed form-meaning mapping test	.911***	1	.978

***p < .001

Not surprisingly, each table for the collapsed scores shows quite high loadings, as in the previous analysis, suggesting that in the scores that are collapsed across the two levels of study time the corresponding test pairs likely measure the same underlying construct.

Figure 18 presents the conceptual structure of the moderated mediation analysis (Model 14) as well as the obtained coefficients in the three factor analytic test scores.

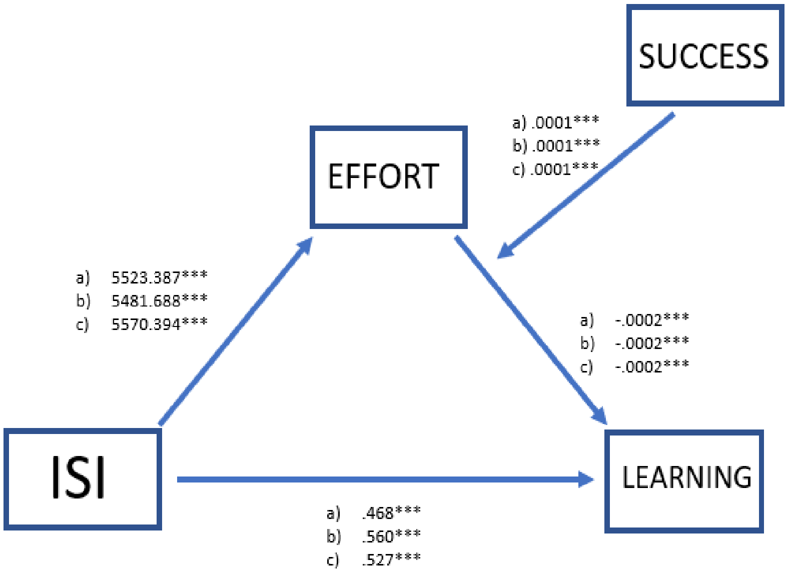


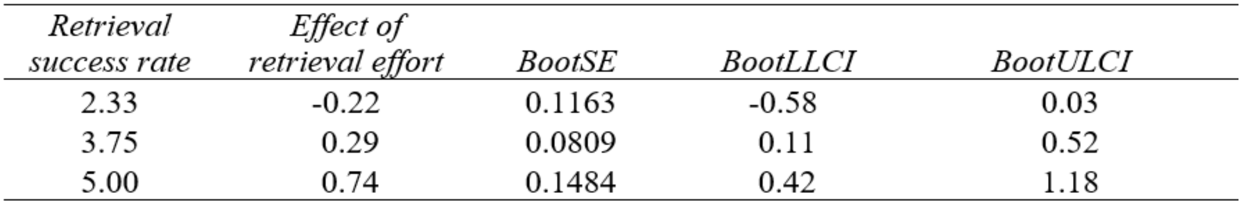
Figure 18: Conceptual structure for the moderated mediation analysis

The coefficients show a similar pattern for all three sets of vocabulary scores. The coefficients show a positive effect of ISI on study-phase retrieval effort and also on the learning outcomes. Effort is shown to actually have a negative effect on learning in each of the three sets of vocabulary scores. Study-phase retrieval success, however, has a positive effect on the relationship between effort and learning.

The form-recognition scores. The test of the indirect effects showed significant moderated mediation, Index of Moderated Mediation = .3579, bootstrapped standard error = .0805, 99% bootstrapped confidence interval [.1935, .5992]. This means that the effect of retrieval effort on form-recognition posttest scores significantly depends on retrieval success.

To investigate more in depth the moderated mediation process, the effect of the mediator was tested at different levels of the moderator variable, in this case, using the 16th, 50th, and 84th percentiles. This analysis is the default in the software used. Table 33 presents the effect of study-phase retrieval effort on form-recognition scores at the three levels of study-phase retrieval success represented by the three percentiles.

Table 33: *Effect of effort at three levels of success for form-recognition*

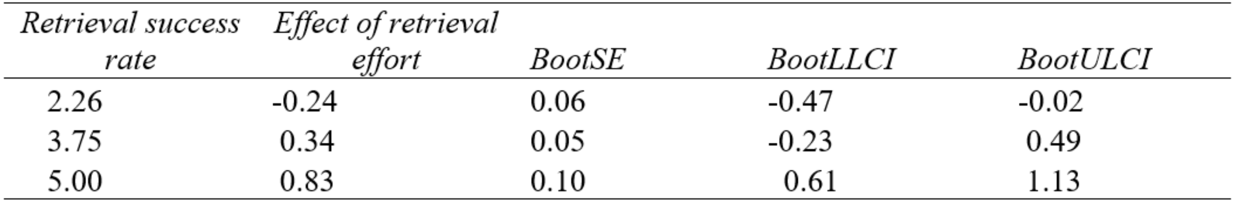


Retrieval success rate	*Effect of retrieval effort*	*BootSE*	*BootLLCI*	*BootULCI*
2.33	-0.22	0.1163	-0.58	0.03
3.75	0.29	0.0809	0.11	0.52
5.00	0.74	0.1484	0.42	1.18

This table shows that effort has a small nonsignificant negative effect for words whose translations were least often successfully retrieved during the study phase (the 16th percentile in retrieval success rate) and a small significant positive effect for words that received an average number of successful retrieval attempts during the study phase (the 50th percentile). For words that received the highest number of successful retrieval attempts (the 84th percentile), however, the effect of effort was larger and significantly positive. Thus, the beneficial effects of effort were shown to be contingent on higher retrieval success in this moderated mediation analysis.

The immediate meaning scores. The test of the indirect effects also showed significant moderated mediation, Index of Moderated Mediation = .3887, bootstrapped standard error = .0588, 99% bootstrapped confidence interval [.2643, .5605]. To investigate more in depth the moderated mediation process, the effect of the mediator was again tested at the 16th, 50th, and 84th percentile levels of the moderator variable. Table 34 presents the effect of study-phase retrieval effort on immediate meaning scores separately at each of the three levels of study-phase retrieval success represented by the three percentiles.

Table 34: *Effect of effort at three levels of success for immediate meaning tests*

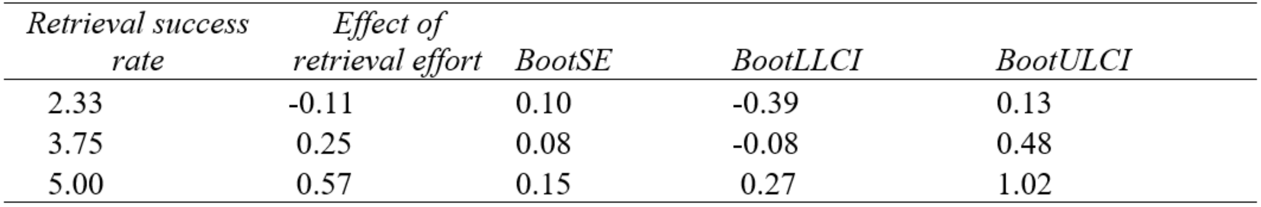


Retrieval success rate	*Effect of retrieval effort*	*BootSE*	*BootLLCI*	*BootULCI*
2.26	-0.24	0.06	-0.47	-0.02
3.75	0.34	0.05	-0.23	0.49
5.00	0.83	0.10	0.61	1.13

A similar pattern is seen for the immediate meaning scores as that for the form-recognition scores discussed earlier, with the exception of a significant negative effect of effort at the lowest level of retrieval success. This latter finding is puzzling because it would suggest that spending more effort on a search of one's memory for the target translation actually hurts memory for the word in question if it is not successfully retrieved. This does not seem to make intuitive sense. One possibility may be item difficulty: a word that a participant has a hard time remembering may lead them to think hard in an effort to retrieve it and still fail to do so. This same word may further be hard to get right on the subsequent test. Thus, my proposed explanation of the obtained pattern of results is not a negative effect of effort on memory but rather the effect of item difficulty on study-phase retrieval effort. The overall pattern, however, is again in line with the predictions of the dual mechanism account under investigation.

The delayed meaning scores. The test of the indirect effects showed significant moderated mediation, Index of Moderated Mediation = .2545, bootstrapped standard error = .0767, 99% bootstrapped confidence interval [.1018, .4860]. This means that the effect of retrieval effort on delayed meaning scores also significantly depends on retrieval success. To investigate more in depth the moderated mediation process, the effect of the mediator was tested at different levels of the moderator variable, in this case, using the 16th, 50th, and 84th percentile. Table 35 presents the effect of study-phase retrieval effort on immediate meaning scores separately at each of the three levels of study-phase retrieval success represented by the three percentiles.

Table 35: *Effect of effort at three levels of success for delayed meaning tests*



Retrieval success rate	*Effect of retrieval effort*	*BootSE*	*BootLLCI*	*BootULCI*
2.33	-0.11	0.10	-0.39	0.13
3.75	0.25	0.08	-0.08	0.48
5.00	0.57	0.15	0.27	1.02

In the delayed meaning scores, similarly to the results of the previous two tests, retrieval effort was shown to only be beneficial with higher levels of retrieval success. However, here retrieval effort is only beneficial at the percentile of success. This is different from the previous two tests, where medium study-phase success percentile also showed a smaller though still significant benefit of effort. Recall that on the delayed meaning tests retrieval effort was shown not to significantly mediate beneficial effects of spacing as a main effect.

The results of the moderated mediation analyses are in line with the predictions of the reminding account, which posits successful effortful retrieval as the mechanism underlying the effects of spacing. At least with regard to overt L2-L1 translation retrieval practice for L2 vocabulary learning in a PAL format, the results show that beneficial effects of effort are conditional on a high level of retrieval success. While a nonmonotonic learning function of lag was not obtained in the posttest scores, the complex underlying relationships included a detrimental effect of spacing that operated through a lower rate of study-phase retrieval success at the longest ISI tested in the present study. This may have affected the magnitude of the lag effect in the present results. In general, while spacing effects are usually found to be large, lag effects are often found to be quite small and inconsistent (Maddox et al., 2018). The fact that the chance of successful retrieval may decrease the longer the lag between repeated encounters or retrieval attempts might be one reason why increases in learning outcomes become smaller the longer the lag.

CHAPTER 5

DISCUSSION

The present research examined the contribution of the dual-mechanism of successful effortful retrieval to the effects of spacing overt L2-L1 translation retrieval practice on learning novel L2 vocabulary in a PAL format with immediate feedback study. It further investigated any effects of the amount of time a learner is given, per encounter and in total, for studying each Finnish word with its translation (presented as feedback after each retrieval attempt) on learning outcomes as well as on the operation of the two mechanisms of retrieval effort and success that are proposed to underlie effects of spacing or lag (Benjamin & Tullis, 2010). Participants (L1 speakers of English) studied 72 novel simple and generic words (half repeating targets and half non-repeating controls) in a language that was completely novel to them (Finnish) in a PAL format, where they were required to attempt to produce the L1 English translation for each L2 Finnish word before being presented with both members of the translation pair for study for either 3 or 9 seconds. The experimental targets repeated based on three repetition schedules: a massed schedule, a short-spaced schedule, and a long-spaced schedule. Participants' study-phase response latency and accuracy were recorded. Three immediate (30-min RI) and delayed (1-2 weeks RI) posttests measured participants' learning gains in terms of form recognition ability and ability to produce and select L1 translations for the 72 studied L2 Finnish words.

The first research question asked whether spacing L2-L1 translation retrieval practice more widely has an effect on learning outcomes as measured by immediate and delayed form-recognition and translation posttests and whether the time (3 vs. 9 seconds) learners are given for study of a Finnish-English translation pair immediately after each retrieval attempt makes a difference for these scores. The results showed a spacing effect of a considerable size across the posttest types and RIs – in other words, the scores for words that were practiced in a massed fashion throughout the study phase were considerably lower than for those in the two spaced practice conditions, regardless of the type or time of test. Importantly, the difference between the massed practice condition and the no-practice control condition was very small, particularly in terms of the long-term gains, where, on the most challenging L2-L1 translation test, scores in the massed practice condition were not significantly different from those in the no-practice condition. Using a no-practice condition in the present study allowed to compare the effects of massed retrieval practice to no retrieval practice as well as to retrieval practice spaced at different intervals. The results suggest that despite the fact that retrieval practice is known to be beneficial for learning, massed practice is not an effective learning tool (sometimes producing learning equivalent to no practice at all), even if it involves retrieval. The present findings are in line with proposals that retrieval from short term memory may not involve processes that make retrieval beneficial for memory (Glover, 1989).

The present study included three levels of lag within the same within-participant experiment. The results showed a significant lag effect (advantage of long-spaced practice over short-spaced practice) on the delayed meaning posttests but not on the immediate meaning posttests, where, on the latter, the longer-spaced condition actually produced slightly less learning than the short-spaced condition (a small non-significant nonmonotonic function of lag). No lag effect (but only a spacing effect) was observed in either of the two form-recognition posttests, where the scores in the short- and long- spaced conditions were very similar. This pattern is in line with previous findings of more pronounced beneficial effects of lag the more challenging the task (Maddox, 2016). Further, the delayed posttests speak to forgetting rates in the different conditions and the present findings are in line with previous psychology research showing that effects of spacing become more pronounced when knowledge is tested after a longer period of time (Bahrick, 1979; Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008; Küpper-Tetzel & Erdfelder, 2012; Rawson & Kintsch, 2005; Rohrer, 2015; Serrano, & Huang, 2018). This suggests that longer spaced practice slows forgetting more effectively than does shorter spaced or massed practice. Thus, despite the fact that retrieval is beneficial for learning, the temporal distribution of retrieval practice may be crucial: massed practice may be not much better than no practice at all and longer intervals between repetitions may produce more robust knowledge that is forgotten more slowly than if the interval between repetitions is shorter. However, see below for the results of the mediation analysis that will show that there is a limit to how widely we can space repeated retrieval or study events before this begins to have a negative effect.

The present results showed that longer study time has a small overall significantly positive effect on the posttest scores, particularly for knowledge of meaning. In the present study, longer study time refers to more time given for the participants to look at and maintenance rehearse the L2-L1 translation pair. Psychology studies have shown that maintenance rehearsal may not be effective for improving memory (e.g., Craik & Watkins, 1873). It is likely true that an important difference between the present findings and those of such psychology studies is that looking at and maintenance rehearsing a novel L2 word form paired with its L1 translation may involve different mechanisms than rehearsing information such as well-known L1 words for a subsequent free recall test, which is the usual learning target in psychology experiments. According to the present results, the time participants are allowed to study a foreign word with its meaning at each presentation in a PAL format might have a small beneficial effect on learning, particularly with spaced practice.

Research has shown that learners are not effective at pacing their own study (Rundus, 1971), often devoting more study time to items that they currently believe to be more difficult, such as to spaced rather than massed repetitions, when this impression may not always be accurate. It was an interesting question whether longer study time that is imposed externally can counteract negative consequences of massing repetitions. The obtained small size of the effect is different, however, from findings from prior SLA research that has shown considerable learning benefits of more attentional processing of L2 words (e.g., Godfroid et al., 2018; Koval, 2019). An important difference may be that in such prior research, learners were free to self-pace their study. This suggests that when longer study time is imposed externally it may not have benefits of the same magnitude as when a participant chooses to devote longer study time to a target word. This, in turn, suggests that the processes that underlie self-regulated and other-imposed longer study time are likely qualitatively different. Recall that the time participants were given for studying a word in the longer study time condition was three times longer than that in the short study time condition. However, the effect of longer study time was dramatically smaller than that of spacing practice, whereby posttest scores resulting from longer study time massed practice were dramatically lower than the scores resulting from shorter study time spaced practice, suggesting that spacing retrieval practice is a more powerful learning tool than externally imposing longer study time.

The second research question asked whether increasing the interval between repeated retrieval events affects study-phase retrieval effort and success, as well as whether the amount of study time that is allowed per encounter affects the relationship between ISI and study-phase retrieval effort and success. The results of the study-phase latency analyses showed that increasing levels of ISI lead to increasing retrieval effort. Retrieval effort decreased slightly from repetition to repetition. The three ISI conditions showed a parallel decrease, with latencies in the long-spaced condition remaining longer than those in the short-spaced condition until the last repetition and latencies in the short-spaced condition, in turn, remaining longer than those in the massed condition until the last repetition. Thus, by increasing ISI, I was able to induce increasingly more retrieval effort, which is known to be beneficial for learning (Roediger & Karpicke, 2006). The amount of time allowed for study of the paired associates per repetition had a very small negative effect on the latencies that was not statistically significant.

The results of study-phase retrieval success analyses showed that retrieval success rate increased with repetition in both spaced conditions, which showed parallel growth with a consistent higher number of successful retrieval attempts across all five repetitions in the short-spaced condition than that in the long-spaced condition. The results showed that (a) in the massed condition, retrieval was almost always successful, (b) in the short-spaced condition retrieval success was significantly less frequent than in the massed condition, (c) in the long-spaced condition retrieval success was significantly less frequent than in the short-spaced condition, indicating that the longer the intervals are between retrieval attempts that are followed by feedback the less successful the retrieval is at respective subsequent retrieval attempts. A growth analysis further showed that there was a small significant positive effect of longer study time in the two spaced conditions but not in the massed condition.

The third research question asked whether the effect of ISI on learning outcomes is mediated by the dual mechanism of successful effortful retrieval and whether the amount of study time allowed per encounter moderates this relationship. The results of moderated mediation analyses showed that the amount of time allowed for study of feedback did not affect the operation of the two proposed underlying mechanisms. They further showed that despite the fact that an overall nonmonotonic function of lag was not obtained in the present learning outcomes, a negative effect of increasing ISI was still present and operated through a lower rate of study-phase retrieval success. Further, it was shown that retrieval success significantly moderated the beneficial effects of more effort that was induced by longer intervals between repetitions on all learning measures used in the present experiment. This confirms the predictions of the dual mechanism of retrieval effort and success that is proposed to underlie effects of spacing on learning by the reminding account of the spacing effect (Benjamin & Tullis, 2010). Recall that, according to this account, retrieval must be effortful yet successful. The present results showed that retrieval effort only had beneficial effects on learning when it was successful.

It is surprising that similar results to those obtained from the L2-L1 translation and form-meaning matching posttests held for the form-recognition tests as well – that is, retrieval effort that is induced through wider spacing of repetitions is only beneficial for form recognition ability when retrieval is mostly successful. While longer retrieval effort should well benefit subsequent ability to recognize target L2 forms due to the fact that longer latencies represent here longer time spent visually processing the L2 form while participants searched their memory for its meaning (Kintsch & van Dijk, 1978), the finding of a benefit of successful retrieval and the finding that longer effort during retrieval attempts was only beneficial when retrieval was successful are quite puzzling. One way that this finding may be explained is that learning is known to be facilitated the more meaningful the stimulus (Marks

& Miller, 1964; Schulman, 1974). It may be that successfully retrieving a meaning associated with an L2 form affects learning of the form because it involves more meaningful processing of the L2 form during the process of retrieval.

Despite the fact that in the present study each retrieval attempt was followed by feedback in the form of the target L2-L1 translation pair, failed retrieval attempts did not benefit from more effort. This is surprising as one would expect a more effortful search of one's memory to result in higher quality processing of subsequently presented feedback, which should, in turn, benefit learning (Izawa, 1970; Kornell, Hays, & Bjork, 2009). Further, there have been proposals that a failed retrieval attempt that is followed by the presentation of feedback in the form of the target searched-for information should have no less learning potential (or even greater learning potential) as does a successful retrieval attempt (Bahrick & Hall 2005; Pashler, Zarow, & Triplett, 2003). The present results showed that lower study-phase retrieval success rate that resulted from spacing retrieval attempts more widely had a negative effect on learning even in the presence of feedback. Further, lower rate of retrieval success that resulted from spacing interfered with beneficial effects of retrieval effort, even in the presence of immediate feedback following each retrieval attempt. This may be due to the fact that failed retrieval attempts that are followed by feedback do not constitute true retrieval events but only constitute input processing that may, nonetheless, be enhanced by the preceding retrieval attempt.

In the present study, learning gains followed a monotonic function of lag, at least for the longer-term gains and for the more challenging tasks of L2-L1 translation and form-meaning matching. This was despite the negative effect of ISI that operated through a lower rate of study-phase retrieval success. One reason for the monotonic function may be the fact that the study-phase task used involves retrieval, which may produce stronger memory traces at each repetition that are more likely to survive longer ISIs (Verkoeijen et al., 2005). This may have prevented dramatic study-phase retrieval failure with the longest ISI used, which in turn failed to have a dramatic negative impact on learning.

Studies investigating effects of equal versus expanding spacing schedules on learning have mostly found an advantage of equally-spaced schedules over expanding schedules (Balota et al., 2006; Carpenter & DeLosh, 2005; Logan & Balota, 2008; Storm et al., 2010). Recall that the main purpose of an expanding schedule is to ensure study-phase retrieval success that can be achieved in this case with progressively longer ISIs. Less learning in the expanding schedules is often attributed to such higher rate of study-phase retrieval success that is promoted through such schedules. Thus, it is argued by some that more failure during study phase may be beneficial. A number of other studies have shown that more study-phase performance failures result in superior learning outcomes (Bahrick & Hall, 2005; Pashler et al., 2003). In the present study, on the surface, the same pattern seems to hold: the long-spaced condition produced the lowest study-phase performance success but learning in this condition was superior in the long term. However, the moderated mediation analyses showed that study-phase performance failure still had a negative effect on learning outcomes. Further, effort put into retrieval attempts that were mostly unsuccessful did not have a positive effect on learning as it should be expected to have for learning from retrieval practice (Bjork, 1975; Glover, 1989; Whitten & Bjork, 1977). The present results suggest that a balance must be struck between study-phase performance success and effort: It appears that the higher effort that is produced by longer ISIs has a powerful beneficial effect on learning providing that retrieval is successful, even when feedback follows the retrieval attempt. Differences in learning gains are therefore likely to be due to the fact that the words that are retrieved successfully though with difficulty are remembered better than those that are not retrieved or are retrieved with minimal effort.

Pedagogical implications

The findings of the present research have important implications for second language vocabulary teaching and learning. The present findings indicate, first of all, that despite the fact that retrieval practice is believed to promote learning in and of itself, how closely together or widely apart retrieval events occur has very important consequences for L2 vocabulary learning outcomes. Using a control condition in the present study allowed me to evaluate the contribution of time spent on retrieval practice under different levels of lag against no practice at all and only a single study event. If retrieval events occur consecutively or in very close succession, such practice may have little to no positive effect on learning, particularly in the long term. Despite the fact that study in the control condition did not involve any true retrieval attempts and only involved one study event that was 3 or 9 seconds in duration, whereas massed practice involved five true (and predominantly successful) retrieval events and six times longer total study of a translation pair, the difference in learning outcomes between these two conditions was very small in the short term and not statistically different from zero on some measures in the long term. This finding suggests that increasing the number of retrieval-restudy events that occur consecutively or closely together (even if this is increased from zero to five retrieval attempts) does not improve learning gains by much and may not be a good way to use study time. Learners are known to often engage in such self-drilling, whereby they repeat a given word with its translation for a considerable length of time, believing that the longer they rehearse it the better it will be remembered; or test themselves on an item that was very recently seen and while retrieval is still very easy because the information still resides in working memory. The present research shows no benefit of such drilling or massed retrieval practice over a single short study event, which, in turn, may produce results that are not significantly different from zero learning gains in terms of long-term retention. To use time effectively, I recommend, therefore, to space retrieval-restudy events. The present results confirm arguments in prior research that spacing practice can help save time: spacing repeated retrieval practice does not require much additional study time beyond the longer time it takes to retrieve the target information; however, it results in far superior learning gains that are more robust to forgetting. With massed practice, it might take much more study to achieve the same learning outcomes (Maddox & Balota, 2015). For learners, I would recommend adopting a more spaced schedule for self-testing and to attempt retrieval of the studied material only once they feel that some, though not complete, forgetting of the target information has occurred. This can be done by interleaving retrieval-restudy of different information rather than using blocked study. Thus, for example, if a learner is studying 20 words with their translations, they may wish to go through the entire list before revisiting any given item rather than devoting a number of consecutive retrieval-restudy events to the same item before moving on to the next item. To use time more efficiently, the learner may also wish to cut study of the same item short as soon as they feel that it has been encoded in memory, without engaging in rehearsal, if the information is to be revisited repeatedly.

Longer intervals between retrieval attempts can be used to enhance learning from retrieval practice and slow forgetting of learned material. The higher retrieval effort that results from longer intervals between repetitions underlies these benefits of more widely spaced retrieval practice. However, the benefit of increased retrieval effort is conditional on retrieval success. This means that, while wider spacing of repeated retrieval is beneficial, retrieval attempts must be scheduled such that retrieval is still mostly successful, which means that retrieval attempts should be spaced but not too widely spaced so that retrieval fails, even when feedback is provided after each such retrieval attempt. The provision of feedback after each retrieval attempt did not cancel out the negative effects of study-phase retrieval failure, suggesting that study-phase retrieval success is important for learning of L2 words with their meaning and its absence cannot be offset by the presentation of the target information as feedback immediately following a retrieval attempt. I recommend that intervals used in retrieval practice, such as those determined by various computer vocabulary learning programs that present words in a format such as PAL or the flashcard method and that use immediate presentation of feedback, need to be spaced rather than massed in order to make retrieval practice more effortful. However, they should not be spaced so widely as to lead to dramatic levels of retrieval failure, as this may cancel out the positive effects of retrieval effort and lead to diminished learning.

When selecting a retrieval practice schedule, we need to take into account the probability of successful retrieval given our specific circumstances and learner variables. Thus, we need to ensure that while increasing intervals between repeated retrieval events produces higher amounts of effort these should not be spaced so widely as to lead to failed retrieval during study, as in such a situation effort may no longer have its positive effects. Many different variables may affect study-phase retrieval success. These may be the difficulty of the studied information (the more difficult it is, the lower the chance of successful retrieval after considerable time), the age group and memory ability of our target population, the complexity and interference potential of the intervening material or activity (which may produce more forgetting, resulting in a lower chance of successful retrieval). Thus, for example, the complexities of more naturalistic contexts, such as those found in classroom learning are likely to decrease the probability of successful study-phase retrieval, interfering with benefits of spacing study more widely (Rogers & Cheung, 2018; Suzuki & DeKeyser, 2017).

Increasing the time, per encounter and in total, that a learner is given to study an L2 word presented with its meaning, such as longer presentation rate in PAL software, has a small overall beneficial effect on memory for a target word and its meaning and also increases the chance of successful retrieval in overt L2-L1 retrieval practice, which was shown in the present study to be important for learning outcomes. Increasing study time does not, however, counteract the negative effects of massing practice, even if such practice involves retrieval. Previous research showed that more attentional processing of the target words leads to more learning (Godfroid et al., 2018; Godfroid, et al., 2013) and may be the reason spacing repeated study results in greatly superior learning outcomes (Koval, 2019; Rundus, 1971). The present results suggest that large benefits of longer study time may be limited to self-regulated learner choice to allocate more attention or effort and may not have benefits of the same size when longer duration is externally imposed on the learner. Therefore, our efforts should be aimed at getting learners to choose to allocate more attention/study time/effort to target forms, such as, for example by using spacing (Koval, 2019) rather than imposing longer study time externally. Computer programs that present immediate feedback after each retrieval attempt need not make feedback presentation longer than is reasonably enough for successful encoding of the information (without additional time to simply rehearse), as doing so appears not to have much benefit and may, therefore, not represent efficient use of study time.

Finally, The results suggest that if there is a chance that a learner may be able to retrieve a given target piece of information from memory, they should be allowed to take the time they need to do so rather than being presented with the information before the retrieval process is complete. It is often tempting, in the interest of time, to present information that a learner might take a long time to retrieve on their own. However, if we rush to present the target information before a learner completes a potentially successful retrieval attempt, this may constitute a less powerful learning event than if the information is fully retrieved from memory.

Limitations and suggestions for future research

The present study has a number of limitations. One of them is the fact that response latencies were measured through a button press, which is not as precise a method as voice-activated recording of latencies, for example. Further, the modality was different between study and test: oral translation was the task during the study phase (and it was timed) but the posttests were in the written modality and participants were given unlimited time to provide their written responses.

The present study investigated the contribution of the dual mechanism of successful effortful retrieval to lag effects in L2 vocabulary learning. Retrieval was operationalized as overt retrieval of the L1 translations for target L2 words in a paired-associate learning format. The results confirmed an important contribution of successful effortful overt L2-L1 translation retrieval to L2 vocabulary learning benefits that come from spacing retrieval practice more widely. It is important to note, however, that overt L2-L1 translation retrieval is only one type of retrieval practice and only one type of retrieval. This type of retrieval is pedagogically interesting primarily because it can be observed. It is an important question whether we need to schedule repeated retrieval events such that they are effortful but still successful in overt retrieval practice, a question that leads to very straightforward pedagogical recommendations. Future research also needs to supplement the present results with an investigation of L1-L2 retrieval practice. Such an investigation is also likely to result in very important pedagogical recommendations that can be applied with relative ease. Based on the findings of the present research, it is quite likely that L1-L2 practice might show a very different pattern in terms of the effect of ISI on learning. The reason for such an expectation is due to the fact that L1-L2 translation, particularly with novel words, is a more challenging task, which is likely to result in dramatically less retrieval success at longer ISIs, which was shown in the present experiment to have a negative effect on learning and also to interfere with beneficial effects of retrieval effort. Shorter ISIs may be found to be more beneficial. Such an investigation may further capture a nonmonotonic function of lag in learning outcomes, which was not observed in the present experiment.

The underlying mechanism of the effects of spaced repeated study of L2 material in a learning situation that does not involve overt retrieval may still depend on a covert retrieval process. Future studies need to explore the contribution of covert retrieval to any effects of spacing study more widely in such learning tasks as well. Such covert retrieval can be observed through tests of simple recognition or through indirect memory tests such as facilitation, or speed-up, in task performance. In Koval (2019), for example, I examined facilitation in reading times on L2 words in my spaced condition with the help of eye-tracking. Significant facilitation was observed in the spaced condition that could not be attributed to simple effects of time. I concluded that such facilitation indicates a study-phase retrieval process in my spaced condition, which likely contributed the considerable beneficial effects of spaced study in my experiment. However, I did not intentionally attempt to vary study-phase retrieval success, but only explored this post hoc. Future studies should attempt to induce study-phase covert retrieval failure through the use of wider spacing to investigate the mechanisms underlying spaced study in learning situations that do not involve overt retrieval.

Although the present study captured negative effects of study-phase retrieval failure, it did not capture a nonmonotonic lag function in learning outcomes. This is most likely due to the fact that while longer ISIs did produce more study-phase retrieval failure, the failure rate was not dramatic. One reason for this may be the fact retrieving an L1 translation for an L2 word is not as difficult a task as L1-L2 translation, for example. Further, because retrieval practice produces stronger memory traces at each repetition, the type of practice used in the present experiment may have further promoted stronger memory traces at each repetition, resulting in a higher rate of study-phase retrieval success. Future research will need to investigate the dual mechanism proposed by the reminding account within a task that may not establish such strong memory traces at each repetition, such as incidental learning of vocabulary from reading comprehension activities (Verkoeijen, et al., 2005). Another reason for not capturing a nonmonotonic function may simply be the fact that the interval between the encounters was not long enough or the intervening activities did not produce enough interference to have sufficient negative impact on study-phase retrieval success and consequently on learning from repeated encounters in this experiment. Future research needs to test the dual mechanism of study-phase retrieval effort and success with longer ISIs.

The present results showed that study-phase retrieval failure had a negative effect on learning outcomes and also cancelled any positive effects of retrieval effort. This is contrary to what has been argued in some proposals in psychology research. Thus, for example, higher rate of study-phase retrieval success is argued to be the reason for expanding schedules producing less learning than do equally-spaced schedules (Bahrick & Hall, 2005; Pashler et al., 2003). Future studies investigating the effects of expanding schedules need to measure effort as well as retrieval success during the study phase in order to be able to make stronger arguments about the complex interplay of retrieval effort and success that may underlie any effects of differentially-intervalled schedules. It may be that the performance success that is supported by such an expanding schedule also has the effect of decreasing effort, counteracting any effects of longer ISIs on study-phase retrieval effort.

The present research investigated the effects of study time at each repetition, which was externally imposed and held at two levels of 3 and 9 seconds. Future research would need to investigate whether the same pattern is observed with activities that involve elaborative rather than maintenance rehearsal (Stoff & Eagle, 1971). Further, study time is only one potentially relevant variable that may affect the operation of the underlying mechanisms of retrieval effort and success. Other relevant variables are numerous. An investigation of their effects on the underlying mechanisms is an important direction for future research. Such research may provide a fuller picture of the conditions under which various amounts of spacing may be beneficial or detrimental for L2 learning outcomes. Further, in the present study, participants studied novel L2 words that represented simple and generic concepts, in a completely novel language, from six repeated L1 translation retrieval attempts that were followed by feedback, within one study session. Future research needs to examine other tasks and learning contexts and other learning targets, as well as other learner proficiencies. It will be important also to test the effects of different numbers of repetitions to explore the effects of relevant variables on the relationship between ISI and learning rate or speed: it may be that fewer repetitions will be needed with spaced practice (Maddox & Balota, 2015) although this may, in turn, depend on other relevant variables and their effects on the mechanisms that underlie learning from different levels of ISI.

APPENDICES

APPENDIX A

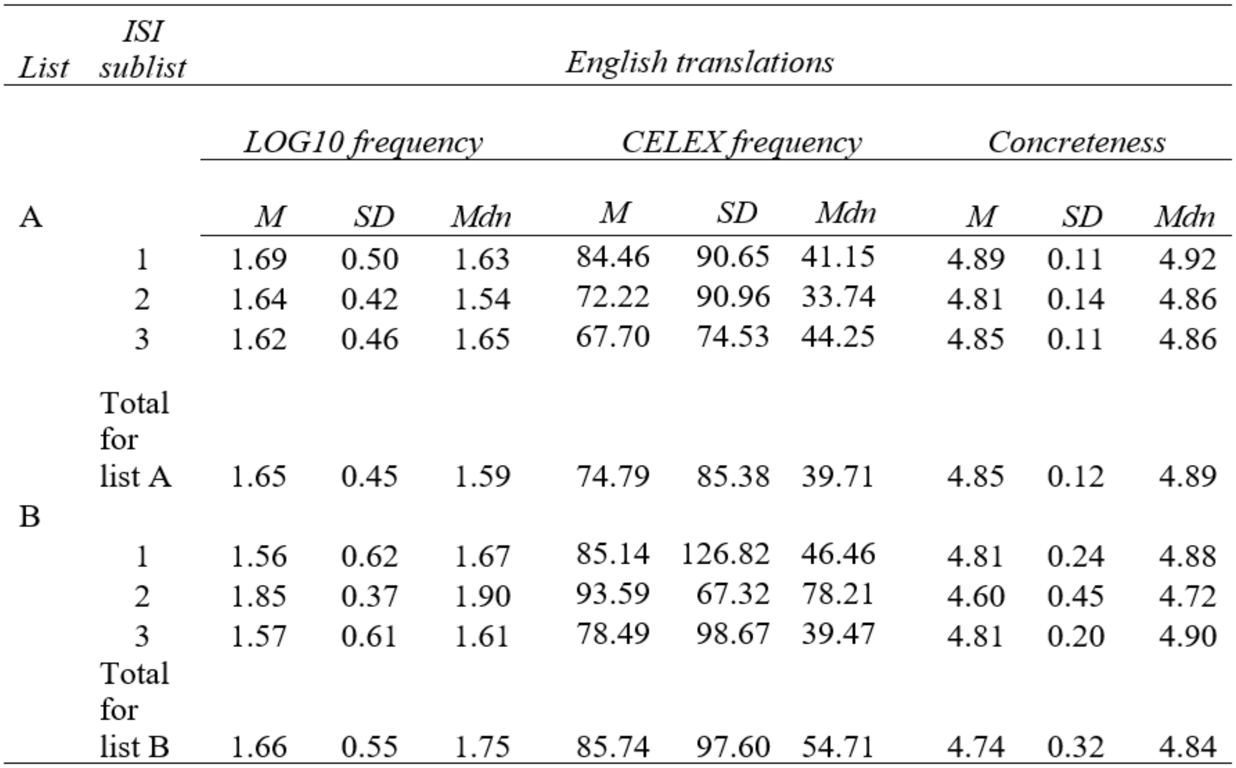
Target Finnish words with their English translations

rakennus	=	building	laukku	=	bag
lehtien	=	leaf	hedelma	=	fruit
sulka	=	feather	perhonen	=	butterfly
sanky	=	bed	ilma	=	air
solmio	=	tie	silta	=	bridge
muna	=	egg	pyrsto	=	tail
pusero	=	shirt	kasine	=	glove
vasara	=	hammer	lapsi	=	child
ruoka	=	food	koira	=	dog
sormi	=	finger	nuoli	=	arrow
lelu	=	toy	piha	=	yard
maaseutu	=	village	hajuvesi	=	perfume
verho	=	curtain	taivas	=	sky
avain	=	key	opettaja	=	teacher
taskuun	=	pocket	kaupunki	=	town
lahja	=	gift	leipa	=	bread
lippu	=	flag	poyta	=	table
orja	=	worker	tarina	=	story
kyna	=	pen	lehma	=	cow
hammas	=	tooth	pilvi	=	cloud
hiekka	=	sand	aurinko	=	sun
keitto	=	soup	jyva	=	grain
ajoneuvo	=	car	veli	=	brother
toimisto	=	office	suihku	=	shower
savuke	=	cigarette	mehu	=	juice
lumi	=	snow	kirjasto	=	library
katu	=	street	kurpitsa	=	pumpkin
kengat	=	shoe	huivi	=	scarf
omena	=	apple	lintu	=	bird
siipi	=	wing	nainen	=	woman
parveke	=	balcony	veitsi	=	knife
kalastaa	=	fish	pelia	=	game
metsa	=	forest	norsu	=	elephant
mekko	=	dress	lusikka	=	spoon
lompakko	=	wallet	lattia	=	floor
tehdas	=	factory	kuva	=	picture

APPENDIX B

Information on the English translations

Table 36: *Frequency and concreteness indices for the English translations for the target words*



List	*ISI sublist*	*English translations*								
		LOG10 frequency			*CELEX frequency*			*Concreteness*		
A		*M*	*SD*	*Mdn*	*M*	*SD*	*Mdn*	*M*	*SD*	*Mdn*
	1	1.69	0.50	1.63	84.46	90.65	41.15	4.89	0.11	4.92
	2	1.64	0.42	1.54	72.22	90.96	33.74	4.81	0.14	4.86
	3	1.62	0.46	1.65	67.70	74.53	44.25	4.85	0.11	4.86
	Total for list A	1.65	0.45	1.59	74.79	85.38	39.71	4.85	0.12	4.89
B										
	1	1.56	0.62	1.67	85.14	126.82	46.46	4.81	0.24	4.88
	2	1.85	0.37	1.90	93.59	67.32	78.21	4.60	0.45	4.72
	3	1.57	0.61	1.61	78.49	98.67	39.47	4.81	0.20	4.90
	Total for list B	1.66	0.55	1.75	85.74	97.60	54.71	4.74	0.32	4.84

APPENDIX C

Instructions for vocabulary posttests

Form-Recognition Test: Please underline the Finnish words that you recognize as ones you have studied during the study phase in this experiment.

L2-L1 Translation Test: Please write the English translation next to each Finnish word below.

Form-meaning Matching Test: Please write the number of each of the English translations below next to its Finnish word on sheet A if you were unable to produce its translation from memory.

APPENDIX D

The form recognition test

Please underline the Finnish words that you recognize as ones you have studied during the study phase in this experiment:

sisavuoren

silta

syyta

sulka

nalka

vasara

akuutin

rakennus

kaveri

lintu

valissa

terve

arvostettu

ihmisen

taskuun

lapsi

jaatelo

ostokset

vuohi

joihin

opetuksen

veitsi

sormi

tilannetta

vihdoin

sianliha

jyva

tikkua

kuumeesta

intohimo

lusikka

piha

nykyinen

tehdas

pelia

verho

taikausko

kuva

pusero

ehka

leijona

perhonen

naen

sana

rasva

kuivempi

vanhasta

lasnaol

ohittaa

lippu

ilma

taso

palvelu

rento

kartano

valmis

pakastin

Appendix D (cont’d)

kyseessa

lompakko

verisia

lehtien

keitto

osamisen

hakijaa

suihku

hiekka

kuten

ampari

kahta

avuton

lattia

neste

kansio

nummi

pojan

opettaja

paistoi

toimisto

taivas

esitys

sanky

puhelin

uhattuna

vahvuus

orja

oikea

lopuksi

hyppasi

samoin

siipi

huulet

etsia

tapahtu

parveke

laukku

nelja

aion

eniten

sitruuna

loyhia

muna

mahtava

eivat

kaiuttim

yllatys

savuke

aurinko

voileipa

hammas

herne

yhdeksan

saimme

metsa

mutkai

loput

luonnolli

maaseutu

ajaksi

muodossa

ongelma

lehma

iloinen

jalka

kasine

haaste

Appendix D (cont’d)

lahja
vihaani
hauska
koyha
mekko
kierrosta
selostus
solmio
ottivat
upea
paras
kertoo
pysyy
lelu
nainen
pyrsto
uhkaa
aasi
katu
hiljaa
epatoivo
erittain
jainen
ajoneuvo
huivi
naytto
kunnes
teltta
alueella
olka
antoi
peto
sohva
leipa
ruoka
pekoni
kehui
summia
tehneet
kalastaa
seka
ankkuria
rotko
riskin
pilvi
kirjasto
harvat
tarve
kutsua
portaat
asettua
hallussa
hedelma
levisi
nuoli
kolme
rullaa
tykonsa
viinissa
takki
otat
voimalle
poyta
veli
uudesta
mehu
anteeksi
tapa

kurpitsa

ikioma

suojaus

omena

menestys

lumi

asteikko

kaupunki

masennus

yskimaan

runous

heista

puhtaus

norsu

vainajan

laastarin

rohkein

etelaisen

uusi

syvenee

tarina

sopeutua

lounas

hajuvesi

olennaisia

koira

nosti

kyna

avain

papu

peruna

varpaat

kengat

sataa

vehna

APPENDIX E

The L2-L1 translation test

Please write the English translation next to each Finnish word below:

taskuun

lompakko

katu

veitsi

omena

pilvi

mekko

ruoka

toimisto

hajuvesi

sanky

ajoneuvo

lelu

lusikka

taivas

kaupunki

pelia

lumi

tehdas

leipa

huivi

silta

nuoli

lattia

savuke

orja

lintu

muna

pusero

kasine

mehu

veli

hedelma

parveke

hammas

opettaja

lapsi

lahja

Appendix E (cont’d)

koira

kengat

norsu

sormi

piha

kurpitsa

jyva

sulka

lehma

ilma

pyrsto

rakennus

keitto

poyta

avain

aurinko

lehtien

kuva

maaseutu

verho

siipi

kalastaa

vasara

suihku

hiekka

kyna

nainen

tarina

kirjasto

metsa

perhonen

laukku

solmio

lippu

APPENDIX F

The form-meaning matching test

Please write the number of each of the English translations below next to its Finnish word on sheet A if you were unable to produce its translation from memory:

1. teacher
2. scarf
3. shirt
4. worker
5. town
6. street
7. table
8. elephant
9. hammer
10. gift
11. sand
12. forest
13. bed
14. apple
15. sky
16. tie
17. curtain
18. building
19. toy
20. wing
21. tooth
22. balcony
23. bag
24. snow
25. grain
26. picture
27. egg
28. wallet
29. pumpkin
30. bread
31. story
32. yard
33. brother
34. bridge
35. knife
36. tail

Appendix F (cont’d)

37. sun
38. game
39. shoe
40. feather
41. key
42. butterfly
43. cow
44. glove
45. car
46. library
47. office
48. perfume
49. cigarette
50. air
51. child
52. fruit
53. soup
54. shower
55. factory
56. arrow
57. food
58. leaf
59. juice
60. pocket
61. dress
62. dog
63. village
64. flag
65. fish
66. woman
67. spoon
68. pen
69. finger
70. bird
71. cloud
72. floor

APPENDIX G

Linguistic Background Questionnaire

Background Questionnaire

1. Participant Number ____________ 2. Gender: M__ F__ 3. Age: _____
4. Native Language(s) ______________
5. Home country or countries: _______________________________________
6. What languages have you studied?

Language	How long have you been studying it?	Age at which you began studying the language	Proficiency Very poor Excellent
			1 2 3 4 5
			1 2 3 4 5
			1 2 3 4 5
			1 2 3 4 5
			1 2 3 4 5

7. Is there anything else you would like to tell us about your language background? If so, please write it here (you can also use the back of this sheet):

8. Did any of the words that you studied in the experiment strike you as familiar upon initial encounter? ___ YES ____ NO
9. If you indicated YES above, please explain:

APPENDIX H

Study-phase instructions

In this experiment, you will study Finnish words with their translations.

You will see words from the Finnish language appear one at a time in the middle of the screen with a question mark prompting you to provide its English translation (jipt -- _____ ?). If you believe that you have studied the word with its translation, please provide the translation by saying it aloud. If you do not believe that you have studied the translation for a given word or if you cannot remember the translation, please say “I don’t know”. Your response time and accuracy will be recorded.

Please try to recall the English translation even if it requires you to think longer.

Please do not try to guess by simply saying translations you have seen if you do not, in fact, consider that it is associated with the word in question.

Sometimes a Finnish word will be presented with its English translation (jipt -- courage). When this happens, please study the word and its translation for as long as it remains on the screen. Please study the Finnish word with its translation each time until it disappears from the screen. Even if you feel that you know the word well while it is still shown, please continue studying it until it disappears.

As soon as a word disappears from the screen and a new word appears, please switch your attention to the new word at once and focus only on the word that is currently being presented.
There will be a test on your knowledge of the Finnish words and their translations after this study phase.

REFERENCES

REFERENCES

Allen, G.A., Mahler, W.A., & Estes, W.K. (1969). Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning & Verbal Behavior, 8,* 463–470.

Arnold, K. M., & McDermott, K. B. (2013). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *39*(3), 940.

Appleton-Knapp, S. L., Bjork, R. A., & Wickens, T. D. (2005). Examining the spacing effect in advertising: Encoding variability, retrieval processes, and their interaction. *Journal of Consumer Research, 32*(2), 266–276.

Atkinson, R.C., & Shiffrin, R.M. (1968). Human memory: A proposed system and its control processes. In K.W. Spence & J.T. Spence (Eds.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 1, pp. 90-195). New York: Academic Press.

Auble, P. M., & Franks, J. J. (1978). The effects of effort toward comprehension on recall. *Memory & Cognition, 6*(1), 20–25.

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1995). The CELEX Lexical Database. Release 2 [CD-ROM]. Linguistic Data Consortium, University of Pennsylvania, Philadelphia.

Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General, 108*(3), 296.

Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science, 4*, 316–321.

Bahrick, H. P., & Phelphs, E. (1987). Retention of Spanish vocabulary over 8 years. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 13*(2), 344.

Balota, D. A., Duchek, J. M., & Paullin, R. (1989). Age-related differences in the impact of spacing, lag and retention interval. *Psychology & Aging, 4*, 3–9.

Balota, D. A., Duchek, J. M., Sergent-Marshall, S. D., & Roediger, H. L., III. (2006). Does expanded retrieval produce benefits over equal-interval spacing? Explorations of spacing effects in healthy aging and early stage Alzheimer's disease. *Psychology & Aging, 21*, 19–31.

Bahrick, H. P., & Hall, L. K. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory & Language, 52*(4), 566–577.

Barcroft, J. (2007). Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning, 57*, 35–56.

Batchelder, W. H., & Riefer, D. M. (1980). Separation of storage and retrieval factors in free recall of clusterable pairs. *Psychological Review, 87*, 375–397.

Begg, I., & Green, C. (1988). Repetition and trace interaction: Super-additivity. *Memory & Cognition, 16*(3), 232–242.

Bellezza, F. S., Winkler, H. B., & Andrasik, F. (1975). Encoding processes and the spacing effect. *Memory & Cognition, 3*(4), 451–457.

Benjamin, A.S., Bjork, R.A., & Schwartz, B.L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127*, 55–68.

Benjamin, A. S., & Ross, B. H. (2010). The causes and consequences of reminding. In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: A Festschrift in honor of Robert A. Bjork* (pp. 71–87). New York: Psychology Press.

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? *Cognitive Psychology, 61*(3), 228–247.

Bird, S. (2010). Effects of distributed practice on the acquisition of second language English syntax. *Applied Psycholinguistics, 31*, 635–650.

Birnbaum, I. M., & Eichner, J. T. (1971). Study versus test trials and long-term retention in free-recall learning. *Journal of Verbal Learning & Verbal Behavior, 10*, 516–521.

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher, & A. Koriat (Eds.), *Attention and performance* (pp. 435–459). Cambridge, MA, US: The MIT Press.

Bjork, R. A. (2013). Desirable difficulties perspective on learning. In H. Pashler (Ed.), *Encyclopedia of the mind* (pp. 242–244). Thousand Oaks, CA: Sage.

Bjork, R. A., & Allen, T. W. (1970). The spacing effect: Consolidation or differential encoding? *Journal of Verbal Learning & Verbal Behavior, 9*(5), 567–572.

Bloom, K.C., & Shuell, T.J. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *The Journal of Educational Research, 74*, 245–248.

Bower, G. H. (1972). Stimulus-sampling theory of encoding variability. *Coding Processes in Human Memory, 3,* 85–123.

Braun, K., & Rubin, D. C. (1998). The spacing effect depends on an encoding deficit, retrieval, and time in working memory: Evidence from once-presented words. *Memory, 6,* 37–65.

Bruce, D., & Weaver, G. E. (1973). Retroactive facilitation in short-term retention of minimally learned paired associates. *Journal of Experimental Psychology, 100*, 9–17.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*(3), 904–911.

Bugelski, B. R. (1962). Presentation time, total time, and mediation in paired-associate learning. *Journal of Experimental Psychology, 63*(4), 409–412.

Bui, D. C., Maddox, G. B., & Balota, D. A. (2013). The roles of working memory and intervening task difficulty in determining the benefits of repetition. *Psychonomic Bulletin & Review, 20*(2), 341–347.

Calkins, M.W. (1894). Association: I. *Psychological Review, 1*, 476– 483.

Callan, D., & Schweighofer, N. (2010). Neural correlates of the spacing effect in explicit verbal semantic encoding support the deficient-processing theory. *Human Brain Mapping, 31*(4), 645–659.

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H., & 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.

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory & Cognition, 19*(5), 619–636.

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test?. *Psychonomic Bulletin & Review, 13*(5), 826–830.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20*(6), 633–642.

Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2009). Optimizing distributed practice: Theoretical analysis and practical implications. *Experimental Psychology, 56*(4), 236–246.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*(3), 354.

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science, 19*(11), 1095–1102.

Challis, B. H. (1993). Spacing effects on cued-memory tests depend on level of processing. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 19,* 389–396.

Collins, L., Halter, R. H., Lightbown, P. M., & Spada, N. (1999). Time and the distribution of time in L2 instruction. *TESOL Quarterly, 33*(4), 655–680.

Commins, S., Cunningham, L., Harvey, D., & Walsh, D. (2003). Massed but not spaced training impairs spatial memory. *Behavioural Brain Research, 139*(1-2), 215–223.

Craik, F.I.M. (1970). The fate of primary memory items in free recall. *Journal of Verbal Learning & Verbal Behavior, 9,* 143–148.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning & Verbal Behavior, 11*(6), 671–684.

Craik, F. I., & Watkins, M. J. (1973). The role of rehearsal in short-term memory. *Journal of Verbal Learning & Verbal Behavior, 12*(6), 599–607.

Crowder, R.G. (1976). *Principles of learning and memory.* Hillsdale, NJ: Erlbaum.

Cuddy, L.J., & Jacoby, L.L. (1982). When forgetting helps memory: An analysis of repetition effects. *Journal of Verbal Learning & Verbal Behavior, 21,* 451–467.

Cull, W.L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology, 14,* 215–235.

Cull, W.L., Shaughnessy, J.J., & Zechmeister, E.B. (1996). Expanding understanding of the expanding-pattern-of-retrieval mnemonic: Toward confidence in applicability. *Journal of Experimental Psychology: Applied, 2,* 365–378.

D'Agostino, P. R., & DeRemer, P. (1973). Repetition effects as a function of rehearsal and encoding variability. *Journal of Verbal Learning & Verbal Behavior, 12*(1), 108–113.

Davis, C. J. (2005). N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods, 37*(1), 65–70.

de Jonge, M., Tabbers, H. K., Pecher, D., & Zeelenberg, R. (2012). The effect of study time distribution on learning and retention: A Goldilocks principle for presentation rate. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 38*(2), 405.

Delaney, P. F., Godbole, N. R., Holden, L. R., & Chang, Y. (2018). Working memory capacity and the spacing effect in cued recall. *Memory*, *26*(6), 784–797.

Deisig, N., Sandoz, J-C,Giurfa, M., Lachnit, H. (2007). The trial-spacing effect on olfactory patterning discriminations in honeybees. *Behavioural Brain Research, 176*(2)*,* 314–322.

Delaney, P.F., Verkoeijen, P., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. In B. H. Ross (Ed.), *Psychology of learning and motivation: Advances in research and theory* (pp. 63–147). San Diego: Elsevier Academic Press Inc.

Dellarosa, D., & Bourne, L. E. (1985). Surface form and the spacing effect. *Memory & Cognition, 13,* 529-537.

Dempster, F.N. (1988). The spacing effect: A case study in the failure to apply the results of psychological research. *American Psychologist, 43*, 627–634.

Dempster, F. N. (1989). Spacing effects and their implications for theory and practice. *Educational Psychology Review, 1,* 309–330.

Dempster, F. N. (1996). Distributing and Managing the Conditions of Encoding and Practice. In E.L. Bjork & R. A. Bjork (Eds.), *Handbook of Perception and Cognition: Memory* (Vol. 10 pp. 317–344). New York: Academic Press.

Donaldson, W. (1971). Output effects in multi-trial free recall. *Journal of Verbal Learning & Verbal Behavior, 10,* 577–585.

Donovan, J. J., & Radosevich, D. J. (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. (1964). *Memory: A contribution to experimental psychology* (H.A. Ruger & C.E. Bussenius, Trans.). New York, NY: Dover. (Original work published 1885)

Elgort, I., & Warren, P. (2014). L2 Vocabulary learning from reading: Explicit and tacit lexical knowledge and the role of learner and item variables. *Language Learning, 64*, 365–414.

Estes, W. K. (1955). Statistical theory of distributional phenomena in learning. *Psychological Review, 62*, 369–377.

Estes, W.K. (1960). Learning theory and the new ‘mental chemistry’. *Psychological Review, 67*, 207–223.

Field, A. (2013). Discovering statistics using IBM SPSS statistics. Sage.

Forster, K. I., & Forster, J. (2003). DMDX: A windows display program with millisecond accuracy. *Behavioral Research Methods, Instruments & Computers, 35*, 116–124.

Gardiner, J.M., Craik, F.I.M., & Bleasdale, F.A. (1973). Retrieval difficulty and subsequent recall. *Memory & Cognition, 1*, 213–216.

Gass, S. (1988). Integrating research areas: A framework for second language studies. *Applied Linguistics, 9*, 198–217.

Gerbier, E., & Toppino, T. C. (2015). The effect of distributed practice: Neuroscience, cognition, and education. *Trends in Neuroscience & Education, 4*(3), 49–59.

Glanzer, M. (1969). Distance between related words in free recall: Trace of the STS. *Journal of Verbal Learning & Verbal Behavior, 8*, 105–111.

Glenberg, A. M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Memory & Cognition, 7*, 95–112.

Glenberg, A. M., & Smith, S. M. (1981). Spacing repetitions and solving problems are not the same. *Journal of Verbal Learning & Verbal Behavior, 20*(1), 110–119.

Glover, J. A. (1989). The" testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81*(3), 392.

Godfroid, A., Ahn, J., Choi, I., Ballard, L., Cui, Y., Johnston, S., ... & Yoon, H. J. (2018). Incidental vocabulary learning in a natural reading context: An eye-tracking study. *Bilingualism: Language & Cognition, 21*(3), 563-584.

Godfroid, A., Boers, F., & Housen, A. (2013). An eye for words: Gauging the role of attention in incidental L2 vocabulary acquisition by means of eye tracking. *Studies in Second Language Acquisition, 35*, 483–517.

Green, J. L., Weston, T., Wiseheart, M., & Rosenbaum, R. S. (2014). Long-term spacing effect benefits in developmental amnesia: Case experiments in rehabilitation. *Neuropsychology, 28*, 685–694.

Greene, R. L. (1989). Spacing effects in memory: Evidence for a two-process account. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 15*(3), 371–377.

Greene, R. L. (1990). Spacing effects on implicit memory tests. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 16*, 1004–1011.

Greeno, J. G. (1964). Paired-associate learning with massed and distributed repetitions of items. *Journal of Experimental Psychology, 67*(3), 286.

Hayes, A. F. (2006). A primer on multilevel modeling. *Human Communication Research, 32*, 385–410.

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis* (2nd ed.). New York: Guilford Press.

Hays, M. J., Kornell, N., & Bjork, R. A. (2013). When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 39*(1), 290.

Hillary, F. G., Schultheis, M. T., Challis, B. H., Millis, S. R., Carnevale, G. J., Galshi, T., & DeLuca, J. (2003). Spacing of repetitions improves learning and memory after moderate and severe TBI. *Journal of Clinical & Experimental Neuropsychology, 25*, 49–58.

Hintzman, D. L. (1974). Theoretical implications of the spacing effect. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola Symposium* (pp. 77–99). Potomac, MD: Erlbaum.

Hintzman, D. L. (1976). Repetition and memory. In G. H. Bower (Ed.), *The psychology of learning and memory* (pp. 47–91). New York: Academic Press.

Hintzman, D. L. (2004). Judgment of frequency versus recognition confidence: Repetition and recursive reminding. *Memory & Cognition, 32*(2), 336–350.

Hintzman, D. L. (2010). How does repetition affect memory? Evidence from judgments of recency. *Memory & Cognition, 38*(1), 102–115.

Hintzman, D. L., Summers, J. J., & Block, R. A. (1975). Spacing judgments as an index of study-phase retrieval. *Journal of Experimental Psychology: Human Learning & Memory, 1*(1), 31.

Hogan, R.M., & Kintsch,W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning & Verbal Behavior, 10*, 562–567.

IBM Corp. Released 2017. IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp.

Izawa, C. (1970). Optimal potentiating effects and forgetting-prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology, 83,* 340–344.

Izawa,C. (1971). The test trial potentiating model. *Journal of Mathematical Psychology, 8,* 200–224.

Izawa, C. (l985b). A test of the differences between anticipation and study-test methods of paired-associate learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 11*, 165–184.

Jacoby, L. L. (1974). The role of mental contiguity in memory: Registration and retrieval effects. *Journal of Memory & Language, 13*(5), 483–496.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning & Verbal Behavior, 17*(6), 649–667.

Jacoby, L. L., Bjork, R. A., & Kelley, C. M. (1994). Illusions of comprehension and competence. In D. Druckman & R. A. Bjork (Eds.), *Learning, remembering, believing: Enhancing team and individual performance* (pp. 57–80). Washington, DC: National Academy Press.

Jacoby, L. L., & Wahlheim, C. N. (2013). On the importance of looking back: The role of recursive remindings in recency judgments and cued recall. *Memory &Cognition, 41*, 625–637.

Johnson, N. F. (1964). The functional relationship between amount learned and frequency vs. rate vs. total time of exposure of verbal materials. *Journal of Verbal Learning & Verbal Behavior, 3*(6), 502–504.

Johnston, W. A., & Uhl, C. N. (1976). The contributions of encoding effort and variability to the spacing effect on free recall. *Journal of Experimental Psychology: Human Learning & Memory, 2*, 153–160.

Joseph, H., Wonnacott, E., Forbes, P., & Nation, K. (2014). Becoming a written word: Eye-movements reveal order of acquisition effects following incidental exposure to new words during silent reading. *Cognition, 133*, 238–248.

Jost, A. (1897). Die Assoziationsfestigkeit in ihrer Abhängigkeit von der Verteilung der Wiederholungen [The strength of associations in their dependence on the distribution of repetitions]. *Zeitschrift für Psychologie und Physiologie der Sinnesorgane, 14*, 436–472.

Kahana, M. J., & Howard, M. W. (2005). Spacing and lag effects in free recall of pure lists. *Psychonomic Bulletin & Review, 12*(1), 159–164.

Kang, S. (2016). Spaced repetition promotes efficient and effective learning: Policy implications for instruction. *Policy Insights from the Behavioral & Brain Sciences, 3*(1), 12–19.

Kang, S., Lindsey, R. V., Mozer, M. C., & Pashler, H. (2014). Retrieval practice over the long term: Should spacing be expanding or equal-interval? *Psychonomic Bulletin & Review, 21*, 1544–1550.

Karpicke, J. D., & Roediger, H. L., III. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval promotes long-term retention. *Journal of Experimental Psychology Learning Memory & Cognition, 33*, 704–719.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319*, 966–968.

Kasprowicz, R. E., Marsden, E., & Sephton, N. (2019). Investigating distribution of practice effects for the learning of foreign language verb morphology in the young learner classroom. *The Modern Language Journal, 103*(3), 580–606.

Kiliç, A., Hoyer, W. J., & Howard, M. W. (2013). Effects of spacing of item repetitions in continuous recognition memory: Does item retrieval difficulty promote item retention in older adults?. *Experimental Aging Research*, *39*(3), 322–341.

Kintsch, W., & van Dijk, T.A. (1978). Toward a model of text comprehension. *Psychological Review, 85,* 363–394.

Kolers, P.A., & Roediger, H.L., III. (1984). Procedures of mind. *Journal of Verbal Learning & Verbal Behavior*, *23*, 425–449.

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*(2), 219–224.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science, 19,* 585–592.

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 35*(4), 989.

Koval, N.G. (2019). Testing the deficient processing account of the spacing effect in L2 vocabulary learning: Evidence from eye-tracking. *Applied Psycholinguistics*, *40,* 1103–1139.

Kuo, T., & Hirshman, E. (1996). Investigations of the testing effect. *American Journal of Psychology, 109*, 451–464.

Küpper-Tetzel, C. E., & Erdfelder, E. (2012). Encoding, maintenance, and retrieval processes in the lag effect: A multinomial processing tree analysis. *Memory, 20,* 37–47.

Landauer, T. K. (1969). Reinforcement as consolidation. *Psychological Review 76*(1), 82–96.

Landauer, T.K., & Bjork, R.A. (1978). Optimum rehearsal patterns and name learning. In M.M. Gruneberg, P.E. Morris, & R.N. Sykes (Eds.), *Practical aspects of memory* (pp. 625–632). London: Academic Press.

Laufer, B., & Hulstijn, J. H. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics, 22,* 1–26.

Linderholm, T., Dobson, J., & Yarbrough, M. B. (2016). The benefit of self-testing and interleaving for synthesizing concepts across multiple physiology texts. *Advances in Physiology Education, 40*(3), 329–334.

Logan, J. M., & Balota, D. A. (2008). Expanded versus equal interval spaced retrieval practice: Exploring different schedules of spacing and retention interval in younger and older adults. *Aging, Neuropsychology, & Cognition, 15*, 257–280.

Maddox, G. B. (2016). Understanding the underlying mechanism of the spacing effect in verbal learning: A case for encoding variability and study-phase retrieval. *Journal of Cognitive Psychology, 28*(6), 684–706.

Maddox, G. B., & Balota, D. A. (2015). Retrieval practice and spacing effects in young and older adults: An examination of the benefits of desirable difficulty. *Memory & Cognition, 43*(5), 760–774.

Maddox, G. B., Balota, D. A., Coane, J. H., & Duchek, J. M. (2011). The role of forgetting rate in producing a benefit of expanded over equal spaced retrieval in young and older adults. *Psychology & Aging, 26*(3), 661.

Maddox, G. B., Pyc, M. A., Kauffman, Z. S., Gatewood, J. D., & Schonhoff, A. M. (2018). Examining the contributions of desirable difficulty and reminding to the spacing effect. *Memory & Cognition, 46*(8), 1376–1388.

Madigan, S. A. (1969). Intraserial repetition and coding processes in free recall. *Journal of Verbal Learning & Verbal Behavior, 8*(6), 828–835.

Mammarella, N., Avons, S. E., & Russo, R. (2004). A short-term perceptual priming account of spacing effects in explicit cued-memory tasks for unfamiliar stimuli. *European Journal of Cognitive Psychology, 16*(3), 387–402.

Marks, L. E., & Miller, G. A. (1964). The role of semantic and syntactic constraints in the memorization of English sentences. *Journal of Verbal Learning & Verbal Behavior, 3*(1), 1–5.

McCabe, D. P., Roediger III, H. L., McDaniel, M. A., Balota, D. A., & Hambrick, D. Z. (2010). The relationship between working memory capacity and executive functioning: Evidence for a common executive attention construct. *Neuropsychology, 24*(2), 222.

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology, 16*, 192–201.

McDaniel, M.A., Friedman, A., & Bourne, L.E. (1978). Remembering the levels of information in words. *Memory & Cognition, 6*, 156–164.

McDaniel, M.A., & Masson, M.E.J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 11*, 371–385.

McKinley, G. L., Ross, B. H., & Benjamin, A. S. (2019). The role of retrieval during study: Evidence of reminding from self-paced study time. *Memory & Cognition*, *47*(5), 877–892.

Medin, D. L., & Schaffer, M. M. (1978). A context theory of classification learning. *Psychological Review, 85*, 207–238.

Miles, S. W. (2014). Spaced vs. massed distribution instruction for L2 grammar learning. *System, 42*, 412–428.

Miles, S., & Kwon, C. J. (2008). Benefits of using CALL vocabulary programs to provide systematic word recycling. *English Teaching, 63*(1), 199–216.

Modigliani, V. (1976). Effects on a later recall by delaying initial recall. *Journal of Experimental Psychology: Human Learning & Memory, 2*, 609–622.

Mohamed, A. A. (2018). Exposure frequency in L2 reading: An eye-movement perspective of incidental vocabulary learning. *Studies in Second Language Acquisition, 40*(2), 269–293.

Melton, A.W. (1970). The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning & Verbal Behavior, 9*, 596–606.

Morris, C.D., Bransford, J.D., & Franks, J.J. (1977). Levels of processing versus transfer-appropriate processing. *Journal of Verbal Learning & Verbal Behavior, 16*, 519–533.

Murdock, B. B. (1960). The immediate retention of unrelated words. *Journal of Experimental Psychology, 60*, 222–234.

Myers, G. C. (1914). Recall in relation to retention. *Journal of Educational Psychology*, *5*, 119–130.

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., & Suzuki, Y. (2019). Effects of massing and spacing on the learning of semantically related and unrelated words. *Studies in Second Language Acquisition, 41*(2), 287–311.

Nakata, T., & Webb, S. (2016). Does studying vocabulary in smaller sets increase learning?: The effects of part and whole learning on second language vocabulary acquisition. *Studies in Second Language Acquisition, 38*(3), 523–552.

Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multi-trial learning of Swahili-English translation equivalents. *Memory, 2*(3), 325–335.

Nelson, T. O., Leonesio, J., Shimamura, A. P., Landwehr, R. F., & Narens, L. (1982). Overlearning and the feeling of knowing. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 8*(4), 279.

Oberauer, K. (2005). Control of the contents of working memory – a comparison of two paradigms and two age groups. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 31*, 714–728.

Pashler, H., Cepeda, N.J., Wixted, J., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, & Cognition, 31*, 3–8.

Pashler, H., Zarow, G., & Triplett, B. (2003). Is temporal spacing of tests helpful even when it inflates error rates?. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 29*(6), 1051.

Pavlik Jr, P. I., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science, 29*(4), 559–586.

Pellicer-Sánchez, A. (2016). Incidental L2 vocabulary acquisition from and while reading: An eye-tracking study. *Studies in Second Language Acquisition, 38*(1), 97−130.

Peterson, L. R., Wampler, R., Kirkpatrick, M., & Saltzman, D. (1963). Effect of spacing presentations on retention of a paired associate over short intervals. *Journal of Experimental Psychology, 66*(2), 206.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory?. *Journal of Memory & Language, 60*(4), 437–447.

Raaijmakers, J. G. (2003). Spacing and repetition effects in human memory: Application of the SAM model. *Cognitive Science, 27*(3), 431–452.

Rawson, K. A., & Kintsch, W. (2005). Rereading effects depend on time of test. *Journal of Educational Psychology, 97*(1), 70.

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition, 14*, 191–201.

Rayner, K., Raney, G. E., & Pollatsek, A. (1995). Eye movements and discourse processing. In R. F. Lorch & E. J. O'Brien (Eds.), *Sources of coherence in reading* (pp. 9–36). Hillsdale, NJ: Lawrence Erlbaum Associates.

Richardson-Klavehn, A., & Bjork, R. A. (1988). Measures of memory. *Annual Review of Psychology, 39*, 475–543.

Robbins, D., & Bray, J. F. (1974). Repetition effects and retroactive facilitation: Immediate and delayed test performance. *Bulletin of the Psychonomic Society, 3*, 347–349.

Robinson, P. (2003). Attention and memory during SLA. In C. Doughty, & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 631–678). Oxford: Blackwell.

Roediger III, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27.

Roediger III, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181–210.

Roediger, H.L., III, & Karpicke, J.D. (2006b). Test enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249–255.

Rogers, J. (2015). Learning second language syntax under massed and distributed conditions. *TESOL Quarterly*, *49*, 857–866.

Rogers, J. (2017). The spacing effect and its relevance to second language acquisition. *Applied Linguistics, 38*, 906–911.

Rohrer, D. (2015). Student instruction should be distributed over long time periods. *Educational Psychology Review, 27*(4), 635–643.

Rohrer, D., & Pashler, H. (2007). Increasing retention without increasing study time. *Current Directions in Psychological Science, 16*(4), 183–186.

Rohrer, D., Taylor, K., Pashler, H., Wixted, J. T., & Cepeda, N. J. (2005). The effect of overlearning on long-term retention. *Applied Cognitive Psychology, 19*(3), 361–374.

Rose, R. J. (1984). Processing time for repetitions and the spacing effect. *Canadian Journal of Psychology/Revue Canadienne De Psychologie, 38*(4), 537–550.

Rose, R. J., & Rowe, E. J. (1976). Effects of orienting task and spacing of repetitions on frequency judgments. *Journal of Experimental Psychology: Human Learning & Memory, 2*(2), 142.

Ross, B. H. (1984). Remindings and their effects in learning a cognitive skill. *Cognitive Psychology, 16*, 371–416.

Ross, B. H., & Bradshaw, G. L. (1994). Encoding effects of remindings. *Memory & Cognition, 22*, 591–605.

Ross, B. H., & Landauer, T. K. (1978). Memory for at least one of two items: Test and failure of several theories of spacing effects. *Journal of Verbal Learning & Verbal Behavior, 17*(6), 669–680.

Ross, B. H., Perkins, S. J., & Tenpenny, P. L. (1990). Reminding-based category learning. *Cognitive Psychology, 22*, 460–492.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432.

Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology, 89*, 63–77.

Rundus, D., & Atkinson, R. C. (1970). Rehearsal processes in free recall: A procedure for direct observation. *Journal of Verbal Learning & Verbal Behavior, 9*(1), 99–105.

Runquist, W.N. (1986). Changes in the rate of forgetting produced by recall tests. *Canadian Journal of Psychology, 40*, 282–289.

Russo, R., & Mammarella, N. (2002). Spacing effects in recognition memory: When meaning matters. *European Journal of Cognitive Psychology, 14*(1), 49–59.

Russo, R., Parkin, A. J., Taylor, S. R., & Wilks, J. (1998). Revising current two-process accounts of spacing effects in memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 24*, 161–172.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*, 29–158.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3–32). New York: Cambridge University Press.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3*(4), 207–218.

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research, 12*(3), 329−363.

Schulman, A. I. (1974). Memory for words recently classified. *Memory & Cognition, 2*(1), 47–52.

Serrano, R. (2011). The time factor in EFL classroom practice. *Language Learning, 61*, 117–145.

Serrano, R., & Huang, H. Y. (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.

Serrano, R., & Muñoz, C. (2007). Same hours, different time distribution: Any difference in EFL? *System, 35,* 305–321.

Shaughnessy, J. J., Zimmerman, J., & Underwood, B. J. (1972). Further evidence on the MP-DP effect in free-recall learning. *Journal of Verbal Learning & Verbal Behavior, 11*(1), 1–12.

Siegel, L. L., & Kahana, M. J. (2014). A retrieved context account of spacing and repetition effects in free recall. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 40*, 755–764.

Soderstrom, N. C., Kerr, T. K., & Bjork, R. A. (2016). The critical importance of retrieval—and spacing—for learning. *Psychological Science, 27*(2), 223–230.

Spitzer, H.F. (1939). Studies in retention. *Journal of Educational Psychology, 30*, 641–656.

Storm, B. C., Bjork, R. A., & Storm, J. C. (2010). Optimizing retrieval as a learning event: When and why expanding retrieval practice enhances long-term retention. *Memory & Cognition, 38*(2), 244–253.

Suzuki, Y. (2017). The optimal distribution of practice for the acquisition of L2 morphology: A conceptual replication and extension. *Language Learning, 67*(3), 512–545.

Suzuki, Y., & DeKeyser, R. (2017). Effects of distributed practice on the proceduralization of morphology. *Language Teaching Research, 21*(2), 166–188.

Suzuki, Y., & Sunada, M. (2020). Dynamic interplay between practice type and practice schedule in a second language: The potential and limits of skill transfer and practice schedule. *Studies in Second Language Acquisition, 42*(1), 169–197.

Thios, S. J., & D'Agostino, P. R. (1976). Effects of repetition as a function of study-phase retrieval. *Journal of Verbal Learning & Verbal Behavior, 15*(5), 529–536.

Thompson, C.P., Wenger, S.K., & Bartling, C.A. (1978). How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology: Human Learning & Memory, 4*, 210–221.

Toppino T. C., & Bloom, L. C. (2002). The spacing effect, free recall, and two-process theory: A closer look. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 28,* 437–444.

Toppino, T. C., & Gracen, T. F. (1985). The lag effect and differential organization theory: Nine failures to replicate. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 11*(1), 185.

Tullis, J. G., Benjamin, A. S., & Ross, B. H. (2014). The reminding effect: Presentation of associates enhances memory for related words in a list. *Journal of Experimental Psychology: General, 143*(4), 1526–1540.

Tullis, J. G., Braverman, M., Ross, B. H., & Benjamin, A. S. (2014). Remindings influence the interpretation of ambiguous stimuli. *Psychonomic Bulletin & Review, 21*, 107–113.

Tulving, E., & Thomson, D. M. (1971). Retrieval processes in recognition memory: Effects of associative context. *Journal of Experimental Psychology, 87*(1), 116.

Van Strien, J. W., Verkoeijen, P. P. J. L., Van der Meer, N., & Franken, I. H. A. (2007). Electrophysiological correlates of word repetition spacing: ERP and induced band power old/new effects with massed and spaced repetitions. *International Journal of Psychophysiology*, 66(3), 205–214.

Verkoeijen, P., & Bouwmeester, S. (2008). Using latent class modeling to detect bimodality in spacing effect data. *Journal of Memory & Language, 59*, 545–555.

Verkoeijen, P., Rikers, R., & Schmidt, H. (2004). Detrimental influence of contextual change on spacing effects in free recall. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 30*(4), 796–800.

Verkoeijen, P., Rikers, R., & Schmidt, H. (2005). Limitations to the spacing effect: Demonstration of an inverted u-shaped relationship between inter-repetition spacing and free recall. *Experimental Psychology, 52*(4), 257–263.

Wahlheim, C. N., Maddox, G. B., & Jacoby, L. L. (2014). The role of reminding in the effects of spaced repetitions on cued recall: Sufficient but not necessary. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 40*(1), 94.

Watkins, M. J., & Kerkar, S. P. (1985). Recall of a twice-presented item without recall of either presentation: Generic memory for events. *Journal of Memory & Language, 24*(6), 666–678.

Waugh, N. C. (1963). Immediate memory as a function of repetition. *Journal of Memory & Language, 2*(1), 107.

Waugh, N. C., & Norman, D. A. (1965). Primary memory. *Psychological Review, 72*(2), 89.

Wenger, S. K., Thompson, C. P., & Bartling, C. A. (1980). Recall facilitates subsequent recognition. *Journal of Experimental Psychology: Human Learning & Memory, 6*, 590–598.

Wheeler, M.A., & Roediger, H.L., III. (1992). Disparate effects of repeated testing: Reconciling Ballard’s (1913) and Bartlett’s (1932) results. *Psychological Science, 3*, 240–245.

Wheeler, M.A., Ewers, M., & Buonanno, J.F. (2003). Different rates of forgetting following study versus test trials. *Memory, 11*, 571– 580.

Whitten, W.B., & Bjork, R.A. (1977). Learning from tests: Effects of spacing. *Journal of Verbal Learning & Verbal Behavior, 16*, 465– 478.

White, J., & Turner, C. (2005). Comparing children's oral ability in two ESL programs. *Canadian Modern Language Review, 61*(4), 491–517.

Xue, G., Mei, L., Chen, C., Lu, Z., Poldrack, R., & Dong, Q. (2011). Spaced learning enhances subsequent recognition memory by reducing neural repetition suppression. *Journal of Cognitive Neuroscience, 23*(7), 1624–1633.

Yin, J. C. P., Del Vecchio, M., Zhou, H., & Tully, T. (1995). CREB as a memory modulator: Induced expression of a dCREB2 activator isoform enhances long-term memory in drosophila. *Cell*, *81*(1), 107–115.

Yonelinas, A. P., & Jacoby, L. L. (2012). The process-dissociation approach two decades later: Convergence, boundary conditions, and new directions. *Memory & Cognition, 40*(5), 663–680.

Young, J. L. (1971). Reinforcement-test intervals in paired-associate learning. *Journal of Mathematical Psychology, 8*(1), 58–81.

Zechmeister, E. B., & Shaughnessy, J. J. (1980). When you know that you know and when you think that you know but you don't. *Bulletin of the Psychonomic Society, 15*(1), 41–44.

Zeelenberg, R., de Jonge, M., Tabbers, H. K., & Pecher, D. (2015). The effect of presentation rate on foreign-language vocabulary learning. *The Quarterly Journal of Experimental Psychology, 68*(6), 1101–1115.

Zimmerman, J. (1975). Free recall after self-paced study: A test of the attention explanation of the spacing effect. *American Journal of Psychology, 88*, 277–291.