The Distributed Practice Effect and Incidental Language Learning

Neil Walker

Submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in Linguistics

Department of Linguistics and English Language

Lancaster University

October 2023

The Distributed Practice Effect and Incidental Language Learning

Abstract

The benefit for spaced compared to massed presentation of to-be-learned items on delayed post-tests, known as the spacing effect, is one of the oldest findings in cognitive psychology. However, despite the robustness of findings in studies investigating distributed practice with paired-associate learning, such as the rote-learning of L2 vocabulary, the findings for studies that have investigated L2 learning under incidental learning conditions are more mixed. Over two studies, I investigated aspects of the temporal distribution of the presentation of L2 grammar and vocabulary when learned under two different incidental learning paradigms. In the first study, I investigated the role that distributed and massed practice play in the learning of an artificial language with nouns, verbs, adjectives and case markers, bound by a verb-final word order under incidental cross-situational learning conditions, and the role that five individual differences in memory (visual and verbal declarative memory, procedural memory, working memory capacity and phonological short-term memory) affected learning and retention. Results from study 1 showed that there was no significant difference in delayed post-test results between massed and distributed practice schedules. However, results suggest that lags may result in a shift in attention to different aspects of the language (from verbs to nouns) for those with strong declarative memory. Building on these findings, in study 2 I investigated whether several factors (intentional vs. incidental learning conditions; items that were presented in training vs. items that require a generalisation of rules; and declarative memory) influence the optimal lag for a 35-day retention interval when learning form-meaning connections (animacy and distance) of four artificial determiners. Results of study 2 mirrored study 1 in that, under incidental conditions, there was no difference between massed and distributed schedules. For the intentional aspect of

the form-meaning connection, distributed practice schedules outperformed massed, with no one optimal lag.

Declaration

I declare that this thesis has not been submitted in the same form for the award of a higher degree elsewhere. I confirm that the work submitted in this thesis is my own, except where work which has formed part of a co-authored publication has been included. My contribution and those of the other authors to the work have been explicitly indicated below. The work presented in Chapter 3 was previously published in the following article:

Walker, N., Monaghan, P., Schoetensack, C., & Rebuschat, P. (2020). Distinctions in the acquisition of vocabulary and grammar: An individual differences approach. *Language Learning*, *70*(S2), 221-254.

The experimental paradigm was conceived and designed by the second and fourth author. This study was conceived by the first second and fourth author. I carried out the review of literature, the data collection and analysis including statistical analyses, and I drafted the paper, while I discussed various issues with the other authors throughout the research project.

31st October 2023

Neil Walker

Acknowledgements

I would firstly like to thank my two supervisors, Patrick Rebuschat and Padraic Monaghan, for their constant support throughout this long, long process. Their advice, support, teaching and encouragement has been invaluable to my PhD and has shown me how to be an academic.

I would also like to thank the following researchers for their advice and feedback on various aspects of the studies that made up this thesis: Michael Ullman, Kara-Morgan-Short, Robert DeKeyser, Jan Hulstijn, Melody Wiseheart, Guillaume Thierry, Morten Christiansen, Aline Godfroid, Raffaella Bottini, Aina Casaponsa, and the anonymous peer reviewers of Language Learning.

I would like to thank Sarah Chadwick for her boundless patience in teaching me just a smidgen of her incredible stats knowhow.

And finally, I would like to thank Laura. Couldn't have done it without you. Wouldn't want to do anything without you. You are the best.

Table of Contents

List of Tables

List of Figures

List of Key Terms

Adj	Adjective
2WFC	Two-way forced choice
ANOVA	Analysis of variance
Aospan	Automated operation span task
Case$_{acc}$	Case marker (accusative)
Case$_{nom}$	Case marker (nominative)
CSL	Cross-situational learning
CVMT	Continuous visual memory test
DT	Delayed test block
ID	Individual difference
ISI	Intersession interval / interstudy interval
GJT	Grammaticality judgment test
L1	First language
L2	Second (or third or more) language
MANOVA	Multivariate analysis of variance
MCM	Multi-scale context model
MLAT-V	Part 5 of the modern language aptitude test
N	Noun
NP	Noun phrase
NRT	Non-word repetition task
NP$_{subj}$	Noun phrase (as subject of the sentence)
NP$_{obj}$	Noun phrase (as objective of the sentence)
OSV	Object-subject-verb
P1, P2, P3	Presentation block 1, 2 and 3
PSTM	Phonological short-term memory
RI	Retention interval

RQ Research question

NRT Non-word repetition task

SLA Second language acquisition

SOV Subject-object-verb

SRT Serial reaction time task

V Verb

VP Verb phrase

Chapter 1: Introduction

1.1 Background to the Thesis

Many a language teacher has wondered how it is that their learners seem to learn a piece of grammar when taught but then have forgotten it by the time the end-of-course test comes around. A strand of second language acquisition (SLA) research focuses on how L2 grammar learning can best be practised and taught to enhance long-term retention. This has led to studies focusing on such questions as whether instruction should be explicit or implicit (Goo et al., 2015; Kang et al., 2019; Norris & Ortega, 2000; Spada & Tomita, 2010), whether practice should be receptive or productive (Shintani, Li & Ellis, 2013) and whether and how feedback should be given (Li, 2010). One area that has only recently gained more interest is that of the temporal distribution of input and practice (see Kim & Webb, 2022 for a meta-analysis). That is, should the presentation and practice of language be bunched all together (massed) or spread out (distributed) over time, and if the latter, what is the optimal spacing to maximise learning?

While L2 vocabulary learned under intentional, paired-associate conditions has consistently been shown to provide a benefit for distributed practice, particularly when tested on a delayed posttest (Bahrick et al., 1993; Bloom & Shuell, 1981; Küpper-Tetzel et al., 2014), research in other areas of L2 learning has provided more mixed results, including L2 vocabulary presented under more incidental conditions (e.g.,Nakata & Elgort, 2021; Webb & Chang, 2015), L2 grammar learning (e.g.,Bird, 2010; Miles, 2014; Rogers, 2015; Suzuki & DeKeyser, 2017a; Kasprowicz et al., 2019), oral fluency and task repetition (e.g.,Bui et al., 2019; and Suzuki & Hanzawa, 2022), and total course hours (e.g., Collins et al., 1999; Serrano & Munoz, 2007). A question remains over whether distributing practice assists in the learning of L2 grammar and vocabulary learned under incidental conditions, and if so what the optimal lag, or gap between

study sessions is. In addition, few studies have investigated the role that individual differences in memory play in L2 learning under massed and distributed schedules.

1.2 Thesis Aims

In this thesis, over two studies I investigated the distribution of language practice under incidental learning conditions. In the first study, I investigated whether an artificial language with nouns, verbs, adjectives and case markers bound by a verb-final word order could be learned under incidental cross-situational learning conditions, what the order of acquisition was, whether learning was durable after 24 hours, and the role that five individual differences in memory (visual and verbal declarative memory, procedural memory, phonological short-term memory and working memory capacity) affected learning and retention.

I also investigated the role that distributed and massed practice play in the learning of this artificial language under such conditions. I investigated whether a 20-minute lag between study sessions improved results on a 1-day delayed posttest compared to massing with no gap between study sessions. I also examined whether the individual difference measures predicted success with the cross-situational learning task under massed or distributed learning conditions and whether there was an interaction between individual difference measures and massed or distributed conditions.

Building on the findings from study 1, in study 2, I investigated firstly whether distributing the exposure of form-meaning connections of artificial determiners that convey animacy and distance (based on the experimental paradigm by Williams, 2005) produce better results on a more educationally relevant 35-day delayed posttest than massing it. I also investigated whether several factors (intentional vs. incidental learning conditions; items that were presented in training vs. items that require a generalisation of the rules; and declarative

memory) influence the optimal lag for a given retention interval when learning form-meaning connections of artificial determiners.

1.3 Thesis Structure

The thesis is organised in the following way: Chapter 2 provides an overall summary of the background literature of the two studies. I first situate the two studies in the field of implicit and incidental language learning and then discuss previous research associated with the two experimental paradigms used in this thesis. I then introduce and define the distributed practice effect, including the lag effect. Next, I discuss the different theories of the underlying mechanisms of the distributed practice effect. I then review the burgeoning literature of L2 distributed practice studies with a particular focus on incidental learning conditions. Finally, I review the small number of studies that have investigated the interaction between distributed practice and individual differences.

The subsequent three chapters (3-5), which describe the two studies in this thesis, are organised in a similar way to a thesis by publication. That is, each chapter has a section that outlines the background literature relevant to each study, methodology, results and discussion. As such, there will be some overlap between the background literature sections in chapters 3, 4 and 5 and the literature review chapter 2.

Chapter 3 and 4 describe, analyse and discuss study 1, in which I investigated the learning of an artificial language under cross-situational learning conditions. Chapter 3 includes the analysis of the order of acquisition of the lexical and syntactic features of the artificial language, the durability of learning after 24 hours, the interrelatedness of learning and the role played by five individual differences in memory (verbal and visual declarative memory, procedural memory, working memory capacity and phonological short-term memory).

In chapter 4, I report on a reanalysis of the data from study 1 taking into consideration massed and distributed practice schedules. I include an analysis of the interactions between distributed and massed schedules and the five individual difference measures.

Chapter 5 describes, analyses and discusses study 2, in which I investigated factors that may influence the optimal lag when learning form-meaning connections for a 35-day delayed posttest.

Chapter 6 includes a general discussion, bringing the findings of the two studies together and discussing implications for our understanding of the distributed practice effect for L2 learning under incidental learning conditions. I also present an update on the skill retention theory by Kim et al. (2013).

Chapter 7 concludes by discussing the contributions of this thesis to theory, methodology and pedagogy. I also suggest some future directions in research into the distributed practice effect and incidental language learning.

Chapter 2: Literature Review

2.1 Structure of the Literature Review

In this literature review, I will first situate the two studies that comprise this thesis within the research area of implicit and explicit learning in SLA, including brief discussions of the literatures associated with the two experimental paradigms used in this thesis. I will then define the distributed practice effect, including the lag effect. Next, I will discuss the underlying mechanisms of the distributed practice effect. I will then review the burgeoning literature of L2 distributed practice studies with a particular focus on incidental learning conditions. Finally, I will review the small number of studies that have investigated the interaction between distributed practice and individual differences. As chapters 3, 4 and 5 include background to the literature sections, there will inevitably be some overlap with the literature review in chapter 2. The rationale is to aid the reader by giving an overview of the literature in this chapter, followed by literature that is specifically relevant to the two studies in chapters 3 to 5, so that the reader is not required to keep returning to this chapter.

2.2 Implicit and Explicit Learning and Knowledge

Teachers, learners, curriculum designers, policy makers, language app designers all have an interest in discovering ways to optimise L2 language learning, and one major avenue of interest lies in research into the relationship between explicit and implicit knowledge (Hulstijn, 2005). This is because many aspects of language learning are thought to rely on implicit knowledge, including comprehension and production in our L1 (N. Ellis & Wulff, 2019), while successive meta-analyses have consistently shown advantages for explicit knowledge (Goo et al., 2015; Norris & Ortega, 2000; Spada & Tomita, 2010). Explicit classroom instruction time is by its nature limited, and therefore, finding ways to manipulate implicit or incidental learning conditions so that learning is optimised outside the classroom is of great importance and lies at

the heart of one strand of SLA research. Both studies in this thesis investigate one such manipulation, that is, the role played by distributing the exposure to grammatical and lexical systems under incidental learning conditions.

Implicit learning is learning which occurs without awareness at the time of encoding (Godfroid, 2022; Reber, 1967) that may result in implicit knowledge, which the learner cannot or at least struggles to verbalise (Williams & Rebuschat, 2022). Implicit learning is often investigated through experiments that include incidental learning conditions, which are those in which participants are not informed about one aspect of the target language item, say the form, and are instead directed towards another aspect of the language item, for example, the meaning. In incidental learning conditions, participants are also not usually told that they will be tested. This may result in either implicit or explicit knowledge. Explicit learning, on the other hand, is learning that is conscious, that is, having an awareness that learning is taking place, and it may result in explicit knowledge, which is verbalizable. Intentional learning conditions often involve informing the participant that they will learn and be tested and giving explicit knowledge about the to-be-learned item. However, Hulstijn (2015) argues that explicit and implicit knowledge and learning should not be seen as dichotomous but rather on a gradient, depending on the level of awareness involved. Berry and Broadbent (2014), writing from a complex systems perspective, agree with a continuum view and suggest that each complex learning task is likely to include a subtle mix of implicit and explicit processes.

How learning takes place under incidental conditions when attention is primarily directed to other aspects of the input has received much interest, including whether learning can take place without awareness (Hama & Leow, 2011; Williams, 2005); the degree to which conscious noticing of form-meaning connections are needed to be made (Schmidt, 1990,1995, 2010;

Tomlin & Villa, 1994); and factors that help shift attention or make items more salient in the input (e.g., the learners' L1, Ellis, 2002; redundancy, Loewen et al., 2009; processing tendencies, VanPatten, 1996). Several of these questions will be discussed at greater length in the following sections.

Researchers have highlighted several low-cost interventions to the learning environment that may increase the learning taking place under incidental learning conditions. These include interventions to make target forms more salient within input and therefore encourage more noticing, such as input flooding and input enhancement (Williams & Evans, 1998); consciousness-raising receptive practice such as in processing instruction (Issa & Morgan-Short, 2019; VanPatten, 2002); encouraging interaction (Mackey, 1999); error feedback during and after meaning-focused productive tasks (Li, 2010; Mackey & Goo, 2007); providing small explicit clues so that learners can discover the rules for themselves (Moranski & Zalbidea, 2022); and the role of sleep in extracting generalities (Batterink et al., 2014). One area in which there have been few studies regarding how to increase noticing in incidental language learning is the temporal spacing of study. Study 1 (see chapters 3 and 4) aimed to shed some light on how the distribution of exposure may interact with individual differences in memory to shift attention.

The two studies in this thesis investigate whether distributing exposure to an unknown language under incidental learning conditions affects the rate of learning and retention after a delayed test. The following two sections give the theoretical background to the experimental paradigms used in the two studies: cross-situational statistical learning (study 1) and implicit learning without awareness (study 2).

2.2.1 Cross-situational Statistical Learning

An increasingly popular area of research into implicit and incidental learning in language as well as other areas of cognition has come from the study of statistical learning (Frost et al., 2019). It is a general-purpose learning mechanism which keeps track of statistical regularities in visual, auditory or tactile input. Statistical learning was demonstrated in eight-month-old infants in a study by Saffran et al. (1996). The infants were able to use statistical information in a stream of speech, namely transitional probabilities between syllables, to make connections between pseudowords and its referent. Statistical learning has been implicated in many aspects of language learning, including word segmentation (e.g., Saffran et al., 1996, Thiessen et al., 2013); phonology (e.g., Maye et al., 2002; Thiessen & Saffran, 2007); and syntax (e.g., Thompson & Newport, 2007). There has been a recent call for implicit learning and statistical learning approaches to be merged and treated as the same phenomena (Christiansen, 2019; Rebuschat & Monaghan, 2019; Monaghan et al., 2019).

Research has extended the findings that children and adults are sensitive to regularities in input to whether learners can implicitly keep track of statistical probabilities across multiple situations to learn words and grammar. Quine (1960) illustrated the difficulty that children face when listening to and trying to make sense of a stream of speech in their L1. If a child sees a scene with a rabbit hopping across a field and hears the word "gavagai" from her parent, the child does not have enough information from this one scene to know whether the word refers to the animal (rabbit), the action (hopping), the feeling (cute) or any number of other possibilities. Cross-situational learning offers a solution to this conundrum Yu and Smith (2007) showed that adults could keep track of cross-situational statistics to learn words and their referents. In their study, they presented participants with slides containing pictures of two, three or four pictures

and they heard the accompanying pseudoword referents. However, they were not told which word referred to which picture. That is, there was not enough information on each trial for the participant to learn the appropriate referent. After as few as six repetitions of each word spread across trials that contained different combinations of words and pictures, a four-way multiple choice test was administered. Results showed that the participants scored significantly above chance. In a follow-up study, Smith & Yu (2008), found that 12 and 14-month-old infants could also rapidly learn word-referent pairs when presented through a series of individually ambiguous situations.

Subsequent research into cross-situational learning has investigated which aspects of language can be learned through multiple ambiguous trials. Studies have shown that nouns (Smith & Yu, 2008), verbs (Scott & Fisher, 2012) and both nouns and verbs simultaneously (Monaghan et al., 2015) can be learned under cross-situational learning conditions. In the study by Monaghan et al. (2015), adult participants saw a dual screen with a moving shape in each scene. Participants heard two words, one referring to one of eight possible shapes and the other to one of eight possible movements. Participants were not told about which words referred to which shape or which action. They had to select the scene they thought the utterance referred to. However, they were not given any feedback on their choice. Results showed that learners could learn both nouns and verbs simultaneously, but with nouns learned more quickly than verbs.

Learning verbs and nouns through cross-situational learning is one thing, but it is still a far cry from naturalistic, language learning situations of a learner's first language or for immersion-like learning contexts for a second language in which learners hear a stream of speech of which the learner may know very little. To extend Quine's (1960) "gavagai" example given above, if the child hears a stream of speech, in order to work out what "gavagai" means based on

what the child is seeing, she must also work out aspects of the grammar, including what part of speech it is, whether it is the subject of the sentence or the object, as well as other grammatical constraints. This problem of how word referents and grammatical categories can be learned simultaneously is a crucial one in language acquisition (Gentner, 1982; Gleitman, 1990; Gleitman et al., 2005). In order to investigate this, Monaghan et al. (2019) extended the experimental design of Monaghan et al. (2015) to include two case markers, *tha* and *noo*, that denoted which word was a noun and which was a verb. In addition to learning the referents for the nouns and verbs, verbal retrospective reports revealed that around half of the participants in the incidental learning group worked out the rules of the case markers, thus showing that grammar and vocabulary could be learned simultaneously through cross-situational statistical learning.

In a series of studies Rebuschat, Monaghan and colleagues further investigated the power of cross-situational learning by extending the number and range of to-be-learned items to closer mimic natural immersion learning by including the referents for words from multiple grammatical categories (Monaghan et al., 2021; Rebuschat et al. 2021; Walker et al., 2020, see chapter 3). Rebuschat et al. (2021) designed an artificial language that included eight nouns (each noun referring to a different cartoon alien), four verbs (referring to actions), two adjectives (referring to colours) and two case markers that denoted the agent and patient of the sentence. The artificial language followed a verb-final SOV or OSV word order similar to Japanese. In experiment 1, participants observed a screen with a dynamic scene, for example, a red alien jumping over a blue alien. Participants simultaneously heard a sentence in an artificial language.

For example:

	haagle	chelad	tha	goorshell	sumbark	noo	fisslin
gloss:	blue	Alien7	OBJECT	red	Alien5	SUBJECT	jumps

Red alien5 jumps over blue alien7

There were 48 unique trials repeated over four training blocks. After each training block, there was a testing block, including 16 test trials. Participants saw a dual screen, in which two dynamic scenes occurred simultaneously side-by-side. The scenes differed depending on which lexical item was being tested. For example, for the noun test trials, the scenes were identical except for the aliens depicted; for the verb test trials, the scenes were identical except for the action. Participants heard a sentence in the alien language and were tasked with choosing which scene the sentence referred. The lexical test trials were then followed by 16 word-order grammaticality judgement test trials, in which participants saw a dynamic scene and heard a sentence in the artificial language. Half the sentences used the grammatically correct SOV or OSV word order; the other half were grammatically incorrect *SVO, *OVS, *VSO or *VOS. Participants needed to decide if the sentence sounded “good” or “funny”. No feedback was given for any of the test trials. Results showed that participants could learn all aspects of the language bar the case markers at a significant above chance score. Verbs and word order was learned first, then nouns, followed by adjectives. Case markers, while not above chance showed evidence of improvement throughout the learning process. In experiment 2, the design was the same except for the training blocks, which became two-scene forced-choice trials. The participant might see, for example, red alien5 jumping over blue alien3 in the left-hand scene and blue alien2 pushing blue alien6 in the right-hand scene. Participants had to decide to which scene the sentence referred. Adding the extra scene was done to add extra ambiguity to the learning

environment, and to allow for further fine-grained charting of learning through the training phase. Results, while slightly lower than in the single-scene experiment 1, were still significantly above chance. In experiment 3, in order to rule out the possibility that the testing blocks interspersed between the training blocks were aiding the learning process, the design was repeated with only one test block at the end of training. Results again indicated that participants were able to learn the artificial language, demonstrating that the testing did not influence the learning. Taken together, these three experiments provided evidence for cross-situational statistical learning of vocabulary and grammar, with no information given about the word referents or their grammatical categories nor any feedback.

Several studies have begun to investigate whether low-cost pedagogic interventions could influence the success of cross-situational learning. In Monaghan et al. (2019), mentioned above, participants were split into an intentional and an incidental group. Those in the intentional group were given instruction regarding the two case markers, *tha* and *noo*. The incidental group were not given any information about the case markers. Results showed that the intentional group outperformed the incidental group but those in the incidental group who became aware of the rules, performed comparably to the intentional group. In addition to explicit instruction, Monaghan et al. (2021) investigated whether another pedagogic intervention, namely explicit feedback, could influence the success of cross-situational learning. The materials and procedure were similar to experiment 2 (dual-scene screen) of the alien artificial language study of Rebuschat et al. (2021) with only a few changes to the methodology. They divided participants into three groups: implicit, explicit and feedback. The explicit group, though not the other two groups, were informed of the two marker words and told that they indicated who the subject and the object of the sentence were. During the training blocks, the feedback group, though not the

other two groups, heard an auditory bell sound when they responded correctly to the choice of two scenes. The feedback group also received feedback during the test blocks, which were given after every training block. Results showed that feedback boosted learning compared to the implicit group, but only for vocabulary and not syntax learning. Interestingly, explicit instruction, unlike Monaghan et al. (2019), did not positively influence learning. Taken together, these studies paint a picture of how cross-situational learning can be boosted by small cost and time-effective pedagogic interventions. These studies may also begin to suggest that this experimental paradigm may be used as a proxy for learning under natural language learning conditions to test the effectiveness of pedagogic and contextual interventions.

One low-cost intervention in cross-situational learning conditions that has so far received very little interest is the temporal distribution of learning under cross-situational learning conditions. In study 1 (see chapter 3 and 4), the experimental paradigm in Rebuschat et al. (2021) is used to investigate the distribution of learning schedules under cross-situational learning conditions, the durability of cross-situational grammar and vocabulary learning after 24 hours and how individual memory differences affect the acquisition of the artificial language and interact with distributed and massed learning schedules.

2.2.2 Implicit Learning without Awareness

Another avenue of research has investigated whether it is possible to learn form-meaning connections without awareness at the level of noticing. A seminal study by Williams (2005) investigating the learning of four pre-nominal determiners (*gi, ro, ul* and *ne*) that signified both animacy and distance suggested that it is possible for form-meaning connections presented under incidental conditions to result in implicit knowledge. Participants were given explicit instruction of the distance aspect of the determiner (intentional condition) but not the animacy aspect

(incidental condition) and were then presented with sentences containing the determiners. While 54% of participants in his study became aware of the incidentally presented mapping, those who couldn't verbalise the rule at the end of the study still managed to score significantly above chance on the posttest. This was evidence, according to Williams, of learning without awareness, or implicit learning. A number of follow-up studies using the same or similar Williams (2005) experimental paradigm have also found evidence for learning without awareness (Batterink et al., 2014; Chen et al., 2011; Kerz et al., 2017; Leung & Williams, 2012; Rebuschat et al., 2013), though other studies using the Williams' (2005) paradigm, or similar, has not always been obtained (Andringa, 2020; Hama & Leow, 2010; Faretta-Stutenberg & Morgan-Short,2011).

Hama and Leow (2010) replicated Williams (2005) study with several modifications to the study design (see Appendix J for a comparison of the study designs). In addition to using a posttest debriefing questionnaire, they used think-aloud protocols as a concurrent measure of awareness. And instead of a two-way forced-choice test, they included a four-way multiple choice test and a productive task. Nine out of 34 became aware of the rule, but for those who remained unaware, animacy scores were only at 53% correct and this was not significant. Hama and Leow concluded that there was no evidence of learning without awareness. However, as Rebuschat et al. (2013) pointed out, it is possible that think-aloud protocols may interfere with the learning process. In addition, the researchers removed 11 of the participants who claimed to be using a non-animacy based strategy for selecting their answers. It is possible that incorrect hypothesis testing does not exclude the possibility that some implicit learning of the correct rule had taken place. Finally, a four-way multiple choice task, rather than providing more fine-

grained data of learning as Hama and Leow claimed, may have increased the chances of a distance-only strategy choice.

In subsequent experiments, Leung and Williams (2012, 2014) modified their methodology to try to better capture any implicit learning that took place. Leung and Williams (2012) reduced the exposure sentences to a determiner plus noun, e.g., gi bull. Participants were shown two pictures on a screen, each with a different object in either a near or far position. Participants then heard the phrase spoken aloud and were asked to indicate as quickly and accurately as possible whether the object referred to was living or non-living. Knowledge of the rules of the form-meaning connections, therefore, helped participants make a quicker decision. And when, in the last test block, participants were presented with determiner plus nouns that violated the rules, a slowdown in reaction times would indicate learning. Results showed that there was a significant slowdown in reaction times for the violation block compared to the exposure blocks even for those participants who reported not becoming aware of the animacy rule. Leung and Williams concluded that this was further evidence of learning without awareness.

Paciorek and Williams (2015), using a different study design, investigated whether the generalisation of form-meaning connections extended to semantic generalisations in collocations. Participants were presented with sentences that contained one of four novel nonsense verbs together with a noun. Two of the nonsense verbs referred to the general meaning of "becoming more of" and two to the general meaning "becoming less of". However, unbeknownst to the participants, one of each of the "becoming more of" and "becoming less of" verbs collocated with concrete nouns (akin to the English words add and deplete) and the other "becoming more of" verb and "becoming less of" verb collocated with abstract nouns (akin to increase and

diminish). Participants were asked to specify whether the verb presented in the sentence referred to "becoming more of" or "becoming less of". After the presentation phase, participants were tested via a false memory task. They were presented with verb-noun collocations, a portion of which had not appeared together in the training phase and were asked to indicate whether they remembered seeing the collocation in the training phase. Results showed that participants erroneously identified as remembering having seen significantly more novel collocations that followed the semantic rule of abstract and concrete nouns than those that broke the rule, even though most did not become explicitly aware of the rule. The researchers concluded that this was evidence of implicit learning of semantic categorisation.

Taken together, these studies appear to show growing, albeit still somewhat tentative, evidence for implicit learning of form-meaning connections, that is, learning without awareness at the level of noticing. One finding shared by all the above studies is that, irrespective of whether there was learning without awareness, there was evidence of incidental learning with awareness. That is, even though attention was directed to another aspect of the language input (meaning), some participants still gained explicit knowledge of the form-meaning connections. One possible future research direction, and the focus of study 2 in this thesis, is to investigate whether learning under incidental conditions, resulting in learning with or without awareness, can be improved by manipulating the learning conditions, and whether such learning is durable at educationally relevant time periods. Study 2 (see chapter 5) investigates the learning of form-meaning connections of the four determiners (*gi*, *ro*, *ul* and *ne*) in Williams (2005) study while distributing learning schedules and testing after 35 days.

2.2.3 Summary

Both of the studies in this thesis investigate the role that the temporal distribution of exposure and practice of an L2 under incidental learning conditions. This section has outlined two avenues of research into incidental learning and the two experimental paradigms used to test them (cross-situational statistical learning and learning without awareness). This section has also investigated the role that small interventions in the learning environment play in helping with L2 language learning. In the next section, one such intervention, namely the temporal distribution of exposure and practice of language will be explored in more detail.

2.3 The Distributed Practice Effect

2.3.1 Definition of the Distributed Practice Effect

The distributed practice effect is an umbrella term for several related phenomena, including the spacing effect and the lag effect. The spacing effect refers to one of the oldest findings in cognitive psychology (Ebbinghaus, 1885/1964) that spacing out the presentation or practice of to-be-learned items (> 0s between blocks or items) confers greater long-term memory retention than massing it (0s), particularly on delayed posttests (Austin, 1921) (see Figure 1). This effect has been found in numerous studies across a number of different domains, including, paired-associate learning (Cepeda et al., 2008), maths puzzles (Rohrer & Taylor 2006, 2007), reading and understanding texts (Rawson & Kintsch, 2005); science-based materials (Reynolds & Glaser, 1964); learning to touch type (Baddeley & Longman, 1978); computer simulation tasks (Shebilske et al., 1999); surgical skill learning (see Cecilio-Fernandes et al., 2018 for an overview). Distributed, or spaced, practice has been shown to benefit the very young and the old (Cornell, 1980; Balota et al., 1989) and even different species (Menzel et al., 2001; Sisti et al.,

2007). Effect sizes range from moderate to large (see reviews and meta-analyses by Donovan & Radosevich, 1999; Delaney et al., 2010; Cepeda et al., 2006; Janiszewski et al., 2003).

Figure 1. The Simplest Distributed Practice Study Design

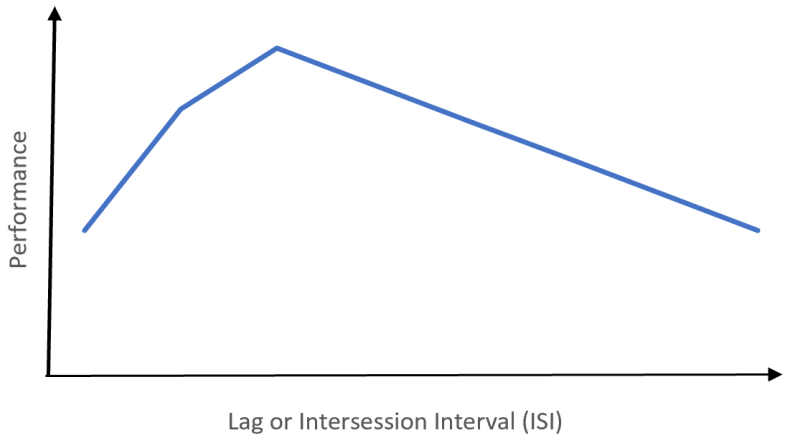
Note. The simplest distributed practice study design with two learning events separated by a temporal lag (or intersession interval, ISI) of greater than 0s. The final test occurs after a delay (retention interval, RI).

The lag effect is the finding that wider gaps tend to confer greater learning effects than narrower ones. More recently, this term has been modified so that it is not just a case of the wider the gap, the better (Küpper -Tetzel & Erdfelder, 2012). Instead, there appears to be a complex non-monotonic relationship between the intersession interval (ISI), the gap between study sessions, and the retention interval (RI), which refers to the gap from the last study session to the delayed test (Küpper -Tetzel, 2014). This relationship resembles an upside-down u-shape, whereby at too narrow or too wide an ISI, performance on a delayed posttest is sub-optimal (see Figure 2). Experimental data suggests that performance is worse for too narrow ISIs than too wide (Gerbier & Toppino, 2015; Rohrer & Pashler, 2007).

An influential study by Cepeda et al. (2008) demonstrated that in addition to this inverted u-shape, the optimal ISI depends on how long you want to remember something for. That is, different RIs require a different ratio of ISI to RI. They asked participants to learn obscure facts over two sessions and tested them with free-recall and multiple-choice tests. They had 26

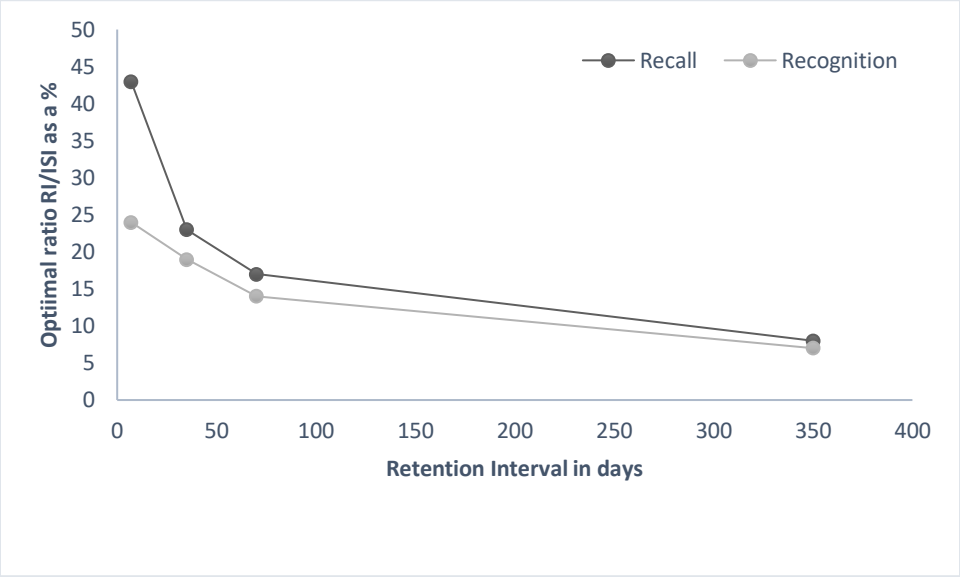
conditions with varying ISIs from 0 to 105 days and RIs from 7 to 350 days. They found that for short RIs, the optimal ISI was around 20-40% of the RI and for the longest RIs, the ratio fell to around 5-10% (see Figure 3). Kim et al. (2019) found similar results after analysing the data from 10,514 individuals who had taken part in online workplace training. They found that longer retention intervals required longer optimal lags. However, a later study by Kornmeier et al. (2014), in which participants studied German-Japanese word pairs and also performed a visual acuity task, found that for RIs of 1 day, 7 days and 28 days, there were two peaks of performance: at 20 minutes and 12 hours. Their study suggested that there was more than one underlying mechanism (see the next section for a summary of different accounts of the underlying mechanisms of the distributed practice effect).

Figure 2. The Lag Effect



Note. The peak represents the optimal spacing. At narrower and wider ISIs, performance is sub optimal. Too wide an ISI tends to be preferable to too narrow.

Figure 3. The Optimal Ratio of ISI to RI of Paired-associate Learning (Cepeda et al., 2008).



Note. Data taken from Cepeda et al. (2008): optimal spacing for intentionally presented paired associates (interesting trivia facts) as a ratio of ISI to RI for recall and recognition tests. At shorter RIs, the ratio of ISI to RI ranges from around 24% for recognition tests to 42% for recall tests. This ratio gradually falls to less than 10% when the RI is around a year.

Finally, often grouped together in the distributed practice family of effects is what Delaney et al. (2010: 65) call the spacing effect's "first cousin": the testing effect. This is the finding that testing is better than restudying, possibly due to the more effortful retrieval (see Roediger & Butler, 2011 for an overview). The testing effect is not a focus of either of the studies in this thesis so will not be referred to again.

As with recent studies (e.g., Küpper-Tetzel, 2014), I will refer to distributed practice to include the spacing effect and the lag effect combined.

2.3.2 The Underlying Mechanisms of the Distributed Practice Effect

Despite hundreds of studies investigating the distributed practice effect over the past century (see Cepeda et al., 2006 for a review), a consensus has yet to be reached regarding the

underlying mechanisms of the distributed practice effect, and a number of possible theories have been put forward to explain them. Any theory of the underlying mechanisms needs to be able to explain the following phenomena and findings from the spacing literature:

1) that spaced practice or presentation is better than massed for delayed posttests but that massed may be better for immediate posttests (e.g., Bloom & Shuell, 1981).
2) the inverted u-shape of performance before and after the optimal ISI (e.g., Cepeda et al., 2008).
3) that the optimal ratio of ISI/RI changes according to the RI (Cepeda et al., 2008)
4) that repetition of spaced presentations increases the chances of successful performance compared to a similar number of independent learning events, that is, there is a dependency between the memory traces, or super-additivity (Maddox, 2016).

In the next section, several different competing accounts of the underlying mechanisms of the distributed practice effect will be discussed.

2.3.2.1 Deficient Processing. According to deficient processing theories (Challis, 1993; Greeno, 1967; Rundus, 1971; Zimmerman, 1975), the amount of attentional processing that is paid to a second learning presentation is reduced when it occurs immediately after the first presentation. Differing versions of deficient processing theory state that this process of reduced attentional resources is either controlled or automatic. In controlled deficient processing accounts, during massed repetitions, a "feeling of knowing" (Callan & Schweighofer (2010; 646) results in less conscious rehearsal taking place (Greene, 1989; Rundus, 1971; Zimmerman, 1975). Rundus (1971) conducted a word learning experiment in which the participants' spoken rehearsal for each item was recorded. He found that with increasing lags between presentations

(in the form of one to seven intervening items), so the amount of rehearsal increased, and so performance on the posttest increased. Zimmerman (1975) designed a study in which participants could choose the amount of time they would spend studying each item in a word-list study that presented items either once, twice or three times. Participants spent less time on repeated items that were presented with massed and short lags (3 items intervening) than with longer lags (14 items intervening). Further, when time spent studying was accounted for, there was no difference between the different lag groups on a free recall test. Zimmerman concluded that these results supported a controlled deficient processing account of the spacing effect.

There are also automatic deficient processing accounts which hypothesise that massing as opposed to spacing presentations of to-be-learned items adversely affects the quantity of processing. In one account, Greeno (1967) suggested that items that are presented one after the other in a massed fashion remain in short-term memory and therefore do not receive the same amount of processing as those items that are spaced. In another account, Challis (1993) proposed that a semantic item that is presented for a second time immediately after the first will be semantically primed and therefore result in less semantic processing. In a spaced learning schedule, on the other hand, the semantic priming will have diminished and therefore more semantic processing will be needed on the second presentation. An eye-tracking study by Koval (2019) that investigated whether deficient processing occurs during the reading of contextually embedded novel L2 vocabulary within sentences, found that participants used less attentional processing for the novel words when they were massed together in consecutive sentences than when they were spaced apart with a 15-20-minute lag. Taken together with the finding that there were statistically higher test scores for the distributed items, Koval concluded that this was in line with deficient processing accounts.

Controlled deficient processing accounts can accommodate learning differences between massed and distributed practice under intentional learning conditions. However, it struggles to reconcile findings that spacing effects still occur in children who do not rehearse (Toppino et al., 2009), and in other species (e.g., Lattal, 1999; Mauelshagen et al., 1998). It is also difficult to see how deficient processing can produce the non-monotonic upside-down u-shape function and the relationship between optimal lag and retention interval, particularly at time periods longer than a few minutes. Nor does it account for super-additivity (Maddox, 2016). Delaney et al. (2010) suggest that deficient processing effects, together with several other effects such as recency effects, are, in fact, imposter effects. Instead of enhancing distributed practice, they reduce the effectiveness of massed practice and should not, therefore, be considered a true distributed practice effect.

2.3.2.2 Study-Phase Retrieval. Study-phase retrieval theories (Appleton-Knapp et al., 2005; Thios & D'Agostino, 1976) posit that studying a to-be-learned item may prompt the retrieval, reactivation and strengthening of the memory of a previous presentation of the to-be-learned item. That is, the first presentation will become more recallable in future. The similarity of the two presentations will also affect the capacity to retrieve the original memory trace with exact repetitions providing a higher likelihood of retrieval than associated materials. Evidence in support of the study-phase retrieval theory came from Madigan (1969), who, in a list learning experiment with a recall test, asked participants to judge how many times a word had been presented during the exposure phase. Only those items that participants had correctly identified as having been presented twice benefitted from the spacing effect. Thios and D'Agostino (1976) provided further support for study-phase retrieval by manipulating the need for retrieval of the

presented items. Only those that required retrieval produced a lag effect. Thus, the study-phase retrieval account depends on the successful retrieval of previously learned items.

However, if being reminded of the previous presentation were the only factor involved in the distributed practice effect, then massed presentations would be superior to spaced ones. Study-phase retrieval theory, therefore, struggles to explain the upward part of the non-monotonic function of performance before the optimal lag. It also cannot adequately explain the ISI/ RI ratio. In order to reconcile these difficulties, later versions of study-phase retrieval theories (see Appleton-Knapp et al., 2005) included a mechanism by which the closer an item is to not being retrieved, the greater it is strengthened. This became part of the reminding account (Benjamin & Tullis, 2010; see below).

2.3.2.3 Contextual Variability. Contextual variability theories (e.g., Glenberg, 1979), also known as encoding variability theories, state that each time an item is encountered, various contextual cues are stored with the item. These can include aspects of the environment (e.g., temperature, location), the learner (mood, tiredness) but also of connections between aspects of the to-be-learned item and other items learned. Final retrieval in the posttest relies on, or is at least affect by, any overlapping cues between the test conditions and the presentation conditions. Therefore, the more different the presentation contexts, the more likely there will be cues that overlap with the testing context and the more likely the item will be retrieved. Glenberg (1979), experiment 1, manipulated the contextual variability of a list of words by presenting them with either the same or a different related word (e.g., for the target word knife, blade-knife in presentation 1 and blade-knife in presentation 2 or blade-knife in presentation 1 and spoon-knife in presentation 2). Participants were then tested on either a cued recall test or a free recall test. Glenberg found that, as per his hypotheses, on the cued recall test no distributed practice effect

was found, due, he believed, to the contextual information already being provided by the cue, and therefore not benefitting from the contextual variability. On the free recall test, on the other hand, in which no contextual cues were provided by the test question, a distributed practice effect was found.

Contextual variability accounts can explain the spacing effect. It can also explain how ISI depends on RI. For shorter RIs, a short ISI would be more likely to overlap with the contextual cues in the final test. For longer RIs, maximising the number of differing cues through spacing out the presentations is more likely to result in an overlap. It struggles, however, to adequately explain the downward slope of the inverted u-shape function of performance, as greater spacing should, according to the theory, continue to increase the chances of contexts overlapping. In addition, several studies have challenged the central premise of contextual variability theories that providing different cues improves recall (Dempster, 1987; Maki & Hasher, 1975). However, see below for theories that have attempted to overcome these difficulties by combining contextual variability mechanisms with other mechanisms in two-factor models.

2.3.2.4 Consolidation and Reconsolidation Theories. Consolidation theories suggest that shorter ISIs do not give enough time for the stabilization and strengthening of memories as longer ISIs (Landauer, 1969). If not enough time is given between the first and the second study session, then the consolidation of the first memory trace will be compromised. An updated version of the consolidation theory of the spacing effect, a reconsolidation account (see Gerbier & Toppino, 2015; Smith & Scarf, 2017), has drawn on more recent research into consolidation processes from a variety of areas. Similar to the original consolidation theory, in reconsolidation accounts, spacing allows time for the initial memory to consolidate. However, rather than the

linear process of memory increase outlined in earlier accounts of consolidation ((Landauer, 1969; Wickelgren, 1972), reconsolidation accounts see the memory trace of the first presentation become initially unstable and malleable upon a second presentation. The second presentation then reconsolidates the first memory trace and makes it stronger. It also appears that reconsolidation processes result in the integration of new learning at a faster rate than if it were learned for the first time (Tse et al., 2007). This theory is supported by research at a cellular level into synaptic plasticity, the functional and structural changes in connection strength between pre and post synaptic neurons that are activated during a learning experience. These changes develop over time, increasing the likelihood of synaptic activation. Research into long-term potentiation (e.g., Bliss & Collingridge, 1993) has shown that the initial simultaneous activation of pre and post synaptic neurons results in a change in the activation structure of the post-synaptic neuron so that henceforth it requires less stimulation from the first neuron. Three phases of long-term potentiation have been differentiated, with decay times in hours, days and months respectively (Abraham, 2003).

Studies into the benefits of sleep on memory have added to this understanding of consolidation processes for lags that are longer than 24-hours. Sleep, and particularly slow-wave sleep in the early stages of a night's sleep, are thought to aid in the consolidation of verbal memory (Ellenbogen et al., 2006; Plihal & Born, 1997, Stickgold & Walker, 2005). Bell et al. (2014), in a study investigating the learning of word pairs, found that sleep consolidation played a role in the distributed practice effect, with a 12-hour ISI (5% ISI/RI ratio) with sleep being just as effective as a 24-hour ISI with sleep (10% ISI/RI ratio) but better than 12 hours without sleep (5% ISI/RI ratio), when tested on a 10-day RI delayed test. Interestingly, studies by Marshall et

al. (2004) and Marshall and Born (2007) suggest that explicitly learned items may benefit more from sleep consolidation than implicitly acquired ones.

Additionally, studies in animals and humans suggest that there is also a qualitative change in the memory after initial consolidation., a process known as systems consolidation (Dudai, 2004). Lehmann et al. (2009) designed a study where rats were given context-shocks in either a spaced (2 shocks per day for six days, 1-day lag) or massed (12 shocks in one day). Then, 7-10 days afterwards, their hippocampus were lesioned. The rats in the spaced group continued to be afraid of the shocks, suggesting their memories were no longer stored in the hippocampus. The rats in the massed group, on the other hand, lost all memory of the context shocks, suggesting that their memories were stored in the hippocampus. In a human word-object pair learning study, Vilberg and Davachi (2013) divided participants into two groups with a second restudy occurring either 20 minutes or 24 hours after the first presentation. Participants' brains were scanned in an fMRI machine as they restudied the word-object pairs. The scans revealed greater connectivity between the hippocampus and the perirhinal cortex for words remembered by the 24-hour lag group than the 20-minute lag group. Taken together, these studies provide strong evidence for neurobiological changes that occur when presentations are distributed compared to massed.

(Re)consolidation theories can adequately explain the spacing effect, although they struggle to account for a massed advantage on an immediate posttest. They can also explain the inverted u-shape function if, as Smith and Scarf (2017) suggest, forgetting is responsible for the downward slope of performance after the optimal lag. According to Smith and Scarf (2017) it can also explain the changing ISI/RI ratio as found in Cepeda et al. (2008), by taking into consideration random effects of item variability (some facts are more memorable than others)

and varying attention given to different facts. Therefore, an optimal lag for any given retention interval, according to Smith and Scarf (2017), must be long enough to benefit some items substantially but short enough so that other items are not too weak.

2.3.2.5 Two-factor and Multi-factor Theories. In an attempt to account for all the distributed practice phenomena mentioned at the beginning of this section, a two-factor model (Verkoeijen et al., 2004) combined contextual variability and study-phase retrieval. According to Verkoeijen et al. as ISI increases, more distinct contextual cues are stored and so increase the chance of retrieval at the final test. However, if the ISI is too long, study-phase retrieval will not retrieve the original memory and contextual cues will instead be stored with a new memory trace. Significantly, the two-factor model accounts for the inverse u-shape of performance.

In a further extension of the two-factor model, Mozer et al. (2009) used computational modelling to produce the Multi-scale Context Model (MCM). MCM adds to the two-factor model a predictive utility component (Staddon et al., 2002), which states that on the second presentation of an item, the length of the preceding gap (ISI) from the first presentation will in part determine how long the memory will be maintained. The shorter the ISI, the less time it will be maintained in memory. Support for this model came from Küpper-Tetzel & Erdfelder (2012), who, using multinomial processing tree analysis on data from a paired-associate task with free and cued recall tests, concluded that the lag effect is the result of encoding and maintenance rather than retrieval processes. Thus, they believe that study-phase retrieval (encoding) and MCM (maintenance) are more likely candidates as underlying mechanisms than contextual variability (retrieval).

Building on the MCM model, Lindsey et al. (2014) developed a computational model adding parameters for item difficulty, learner ability, past study history and forgetting. These

parameters result in a personalised, item-specific spacing schedule rather than one optimal ISI for each given RI. When tested on middle school Spanish learners with a computer-based vocabulary review programme, spacing schedules designed according to this model fared better than MCM and other models. The focus of experiment 2 of the current study is to break down those parameters and determine what aspects of learner ability and item difficulty affect optimal ISI.

An alternative two-factor theory that tries to explain all of the phenomena is the reminding account (Benjamin & Tullis, 2010). This builds on study-phase retrieval accounts by combining Hintzman's (2004) reminding mechanism with Bjork's (1994) desirable difficulties theory. In the reminding account, as in study-phase retrieval accounts, memory traces are strengthened on the second presentation of a to-be-learned item when the learner is reminded of the first presentation. The potential for reminding varies from high capacity for repetitions of the first presentation, medium capacity for associates or variants of the first presentation, and low capacity for items unrelated to the first presentation (Benjamin & Tullis, 2010). The degree to which a memory is strengthened by reminding depends on the difficulty involved in retrieving it, with higher levels of difficulty arising from either a large amount of forgetting (from a wide lag) or a low amount of reminding (from unrelated or associates). The optimal lag will be one in which a desirable degree of difficulty is involved. If the difficulty is too great, with too wide a lag, learners will not be reminded of the first presentation. If, on the other hand, the difficulty level is too low, with too narrow a lag, while the second presentation will remind the learner of the first, the amount of strengthening that occurs will be less. Similarly, in this account, repetitions will require a wider optimal lag than associates. While Benjamin and Tullis (2010) conceptualised reminding as an automatic process, studies by Wahlheim et al. (2014) and Bui et

al. (2014) suggest that reminding can be brought under conscious control and this can enhance the spacing effect.

Desirable difficulty has been operationalised in several different ways. One way is through learning phase accuracy, whether an item can be successful retrieved. Jacoby (1974) measured learning phrase accuracy by using a category recognition structure to sidestep potential confounds with item difficulty. Participants were asked to state whether the item being presented was of the same semantic category as the previous item or previous n items. Jacoby found a role for reminding difficulty as measured by the learning phase accuracy. Another proxy for desirable difficulty is response latency, and it has been used in a number of studies (e.g.,Karpicke & Roediger, 2007; Logan & Balota, 2008; Maddox et al. 2018). Response latency and accuracy of a recognition detection task were used in a series of experiments by Maddox et al. (2018) to investigate whether desirable difficulty or encoding variability affected final memory performance for spaced repetitions within a list-learning paradigm. Participants were presented with a list of words presented one or two times separated by one or five items. Participants were asked to judge whether an item had been previously presented, and to do so as quickly and accurately as possible. In order to directly compare desirable difficulty and encoding variability, in experiment 3 an additional encoding variability variable was added by presenting the second, repeated item in the voice of either the same gender or a different gender. Results showed that while difficulty remained the same for different lags, there was a significant difference in the final test, thus suggesting that the desirable difficulty mechanism did not play a role in the lag effect. Encoding variability, on the other hand, in the form of presenting the items in different gendered voices, was found in experiment 3 to affect the lag effect.

The reminding account explains the inverted u-shape function of test performance compared to ISI. As ISI increases, the difficulty associated with retrieving the item likewise increases, accounting for the upward section of the inverted u-shape function. However, at the same time, as ISI increases the likelihood of successful reminding will decrease, which explains the downward portion of the u-shape when the difficulty becomes undesirably effortful and memory fails. One aspect of distributed practice research findings that the reminding account struggles to explain is, according to Maddox (2016), the finding that massed practice schedules often produce better results on immediate posttests than spaced schedules. If an item has a short RI, reminding accounts would predict that retrieval would be more desirably difficult in the spaced schedule than the massed. Thus, while the reminding account explains a good deal of the findings of distributed practice research, it may not be enough to fully explain all the findings.

2.3.2.6 Summary of Underlying Mechanisms. In summary, there is still no consensus as to the underlying mechanism or mechanisms of the distributed practice effects. In order to adequately explain the consistent findings outlined above, the more recent trend for dual and even multi-mechanism accounts may be a more accurate explanation than single-mechanism accounts. It is also possible the same mechanisms do not underlie the distributed practice effect under all conditions. For example, while deficient processing accounts (both controlled and automatic) may play a role in lags ranging from massed to a few seconds, as in many list learning experiments, it is unlikely to play much, if any, of a role in lags in the hours or days. And, if more than one underlying mechanism does contribute to the distributed practice effect, then other factors, such as type of task, may also influence the relative importance played by particular mechanisms.

2.4 The Distributed Practice Effect and Second Language Acquisition

Learning an L2, whether via classroom study, self-study or through immersion conditions, requires a huge commitment of time and effort. Therefore, finding the most time-efficient learning schedule is of clear interest to educators, curriculum designers, app designers and learners. Findings from the wide range of domains that benefit from distributed practice offer the potential for assisting language learning. However, several factors suggest caution needs to be exercised before generalising the findings from cognitive psychology to L2 acquisition. Firstly, while distributed practice effects have been demonstrated in a wide range of domains, a large number of studies were carried out with paired-associate learning tasks (e.g., Bloom & Schuell, 1981; Cepeda et al., 2008; Küpper-Tetzel et al., 2014). While there is a direct relevance to L2 vocabulary learning under rote-learning (paired associates learning) conditions, in, for example, computer-based flash card systems (see Nakata, 2015 for an overview), there is considerable uncertainty over whether this translates to either vocabulary learning under more incidental, naturalistic conditions or L2 grammar learning.

Secondly, many distributed practice effect studies train participants to mastery on the first block and then merely review on subsequent blocks (e.g., Lindsey et al., 2014; Cepeda et al., 2008). L2 grammar acquisition, in particular, consists of a slow accumulation of understanding and skill with mastery rarely if ever achieved in the first block of practice or presentation (Paradis, 2009). It is unclear, therefore, whether distributed practice will benefit all aspects of L2 language learning in the same way. In addition, when spaced out, L2 grammar may be more likely to be disrupted by reconsolidation than vocabulary items learned to mastery in a paired-associates spacing paradigm. When memories are retrieved, they become unstable and are more

liable to be distorted (Gerbier and Toppino, 2015), which arguably is more of an issue when the to-be-learned item has not yet been mastered.

Finally, many spacing studies were carried out with ISIs and RIs in the seconds and minutes rather than in days, weeks and months, which are more relevant units of time from an educational perspective. It has been suggested that some of the contributory factors at those small timescales may not apply at longer ones. For example, the deficient processing account (e.g., Hintzman, 1974) states that if an item is still in working memory, less attention will be paid to it on the second presentation. This clearly becomes less relevant when ISIs reach hours and days. In addition, paradigms that use ISIs of seconds are nearly all list learning, where rather than being distributed by temporal gaps consisting of no study, to-be-learned items are interleaved with other to-be-learned items. Several studies have found evidence that interleaving is a separate phenomenon from temporal spacing particularly for category learning (see Guzman-Munoz, 2017 for a summary).

The above factors may account for some of the mixed findings into whether L2 acquisition benefits from distributed practice. What follows is a more detailed review of distributed practice studies into different aspects of L2 learning.

2.4.1 L2 Vocabulary Learning

2.4.1.1 Paired-Associate L2 Vocabulary Learning. The one area of L2 learning for which there is no doubt that distributed practice benefits is L2 vocabulary learning under rote-learning (or paired-associate learning) conditions (Bahrick et al., 1993; Bloom & Schuell, 1981; Küpper-Tetzel et al., 2014; Gerbier et al., 2015; Kornmeier et al., 2014). In one example of a longitudinal spacing study, Bahrick et al. (1993) investigated the learning of 300 English-foreign language (French for 3 participants and German for the other) word-pairs over either 13 or 26 learning sessions with ISIs of 14, 28 or 56 days. Participants were presented with and then tested to mastery on all 300 words in each session. Word-pairs were divided into 50-word groups and tested at either 1, 2, 3 or 5 years after the last training session using free recall. Results showed that word-pairs learned in the 56-day ISI schedule were recalled at a significantly higher rate than narrower lags on the 5-year RI (ISI/RI ratio of 3% for 56-day ISI compared to >1% for the 14-day ISI). Another rote-learning vocabulary study by Bloom and Shuell (1981) demonstrates how the distributed practice effect often only appears on delayed posttests. They taught 56 American high school students 20 French-English word pairs in three sessions scheduled in either a massed (3 x 10-minute session in one day) or distributed (1 x 10 minute per day for 3 days). Recall tests were administered on an immediate posttest and four days later (ISI/RI ratio of 25% for the distributed group). Results showed that while there was no difference on test scores on the immediate posttest, there was a significant advantage for the distributed group on the 4-day delayed posttest.

2.4.1.2 Incidental and Contextual Learning of L2 Vocabulary. While many studies have found evidence for the benefit of distributing L2 paired-associate vocabulary learning, the picture becomes less clear with other aspects of L2 learning, including L2 vocabulary learned

under more naturalistic, incidental learning conditions. In recent years, there have been a number of studies that have investigated the distribution of vocabulary learning during extensive reading (Elgort & Warren, 2014; Macis et al., 2021; Serrano & Huang, 2018; Webb & Chang, 2015), but a consensus has yet to be reached as to whether incidental learning of vocabulary benefits from distributed practice. A contextual vocabulary learning study by Koval (2019) compared massed and distributed (ISI, M=18.2 minutes) learning of 24 novel (Finnish) words within sentence contexts, which were then tested on a 24-48-hour delayed posttest (ISI/RI ratio of 1.3% for the distributed condition). The study used eye tracking to determine whether the participants devoted more attentional resources to the novel words under distributed conditions than massed. She found that there was more attention paid to the novel words that were distributed and that there was a clear distributed practice effect. However, as Nakata and Elgort (2021) point out, as the participants were instructed to learn the words a for a posttest, this study might be better described as using intentional rather than incidental learning conditions.

Another study also found an advantage for distributing vocabulary when learned in context. Serrano and Huang (2018) gave Taiwanese high school students a text to read five times, either once a day for five days or once a week for five weeks. The text was embedded with 36 target vocabulary items. \While reading, participants were encouraged to read for comprehension rather than focusing on the unknown words. However, after reading, participants were given a glossary with the meanings of the target words. A bilingual vocabulary matching task was administered both on an immediate posttest and on a delayed posttest (28-day RI for the 5-day lag group, ISI/RI ratio of 25%; 4-day RI for the 1-day lag group, ISI/RI ratio of 25%). Results revealed that participants in the 1-day lag group scored more highly on the immediate posttest than the 5-day lag group, but on their respective delayed posttests, the 5-day lag group

outperformed the 1-day lag group. The design of the differing retention intervals leaves a question over what the optimal lag is for each group. In addition, the use of a glossary to give the participants the definitions of the target words arguably makes this experiment less incidental, in that the participants did not have to infer the meaning themselves. It is possible, therefore, that the benefits of the longer lag group might not be the same if the participants were not given a glossary.

Looking at incidental vocabulary learning from a different angle, Vlach et al. (2012) used a cross-situational learning paradigm to investigate vocabulary learning under more immersive, incidental conditions. In cross-situational learning conditions participants are not given enough information about the vocabulary on each trial to make form-meaning mappings, but over a number of trials, a build-up of statistical probabilities and hypothesis testing (see Yurovsky & Frank, 2015 or the previous section, 2.2.1, for an overview of cross-situational learning paradigms) allows participants to learn the vocabulary. Vlach et al. (2012) found that two-year olds were better able to abstract and generalise nonce nouns under cross-situational learning conditions when schedules were distributed compared to massed. However, ISIs were in seconds and the RI was at 15mins, and only one noun was learned at a time. It has yet to be determined if distributing practice over longer periods and with multiple lexical categories confers a similar benefit and whether this translates to adults exposed to an L2.

Other studies have found no advantage for distributed practice of contextual vocabulary studies (Elgort & Warren,2014; Webb and Chang, 2015). Elgort and Warren (2014) investigated the role of spacing in incidental learning of vocabulary while reading as part of a study that investigated a number of different factors that might influence the acquisition of incidental vocabulary learning. The authors reported that the 48 adult participants mostly, but not always,

read a chapter within a day and subsequent chapters after a lag of one or two days (four chapters in ten days). In contrast to Serrano and Huang (2018), they found that encountering a word (in their case pseudo-words) multiple times in the same chapter resulted in greater retention on an immediate posttest (exact RI not reported but assumed to be one-day; ISI/RI ratio of 100% for across-chapters), using a meaning generation task. However, without a delayed posttest with a retention interval somewhere around the optimal ratio of 10-30%, it is impossible to draw too many comparisons with the other incidental vocabulary studies. Webb and Chang (2015), in a contextual vocabulary learning study involving Taiwanese high school students, found no effect for the distribution of vocabulary, when encountered across ten graded readers.

In a similar contextual vocabulary study by Nakata and Elgort (2021), 48 vocabulary items were presented three times each in sentence contexts in either a massed (consecutive sentences) or spaced (25-minute lag) schedule. They then administered delayed posttests (2-day lag, ISI/RI ratio of 8.7%) involving both more explicit meaning recall and meaning-form matching task and a more implicit priming task. They found that there was a spacing effect for the more explicit meaning recall and meaning-form matching tasks but not for the more implicit priming task, suggesting that implicit knowledge and more tacit semantic knowledge may not benefit as much from distributed practice as the development of explicit knowledge.

Finally, in a recent study, Macis et al. (2021) investigated the learning of L2 collocations under incidental and deliberate learning conditions. In experiment 1 participants were given one text to read a week for five weeks that either contained one collocation embedded five times in the text (massed group) or 25 collocations embedded once in the text (spaced group). Participants' attention was not drawn to the collocations; instead, they completed comprehension questions (not involving the collocations) about the text. Participants were then given a delayed

posttest in the form of a cued form recall task three weeks after the last treatment session (ISI/RI ratio of 33% for the distributed group. In experiment 2, with the same procedure, lags and retention interval, participants were asked to deliberately study collocations in lines of concordance. Results revealed that spacing had a large effect in the deliberate learning condition and a small effect in the incidental learning condition. However, they also revealed that massing under the incidental learning condition produced significantly higher test scores than spacing. They hypothesised that the difference in test results may lie in the amount of noticing that took place. Massing the target collocations within one text made them more salient and more likely to be noticed than in the spaced incidental condition (Schmidt, 1990, 2012). For the deliberate practice condition, as participants were asked to focus on the underlined target collocations in short concordance lines, there was a much greater chance of noticing, even in spaced conditions. And from a reminding account perspective (Benjamin & Tullis, 2010), the spaced repetitions in the incidental condition would be at a more desirable level of difficulty.

Taken together, the recent batch of incidental vocabulary learning spacing studies offer somewhat conflicting evidence, but it is certainly less clear that there is a distributed practice effect than for paired-associate learning tasks. It is possible that vocabulary learned in more deliberate, intentional ways benefit more from distributed practice as the shorter spaced or even massed conditions may be at a desirable level of difficulty for the incidentally encountered items. In addition, as Macis et al. (2021) suggest, the more massed conditions may make incidentally presented vocabulary more salient and therefore induction more likely.

2.4.2 Distribution of Course Hours

Another area of SLA research in which the effects of distributed practice are less clear is the distribution of course hours (Collins et al., 1999; Serrano & Munoz, 2007). These studies are

of particular interest to language teachers as they evaluated the effectiveness of intensive versus part-time courses, a relevant choice that many institutions and students have to make. Serrano and Munoz (2007) compared 110 hours of English language instruction over either 7 months (4 hours per week), 3 to 4 months (8-10 hours a week) or 5 weeks (25 hours a week). In vocabulary, grammar, reading and listening tests, students performed better in the more intensive groups. Collins et al. (1999) compared 350-400 hours of English language instruction over either 5 months or 10 months. Tests included vocabulary recognition tasks and oral narration tasks. Those in the more intensive course performed better than those in the longer course. Studies such as these cast doubt on the effectiveness of distributed practice at more educationally relevant time schedules. However, as Rohrer (2015) pointed out, several issues with the study design, mean caution should be exercised when comparing these results with those from cognitive psychology and later spacing studies. Firstly, neither of these studies included delayed posttests. It would be interesting to find out whether the advantage for the more intensive courses were maintained at a delay posttest after a month or more. Secondly, in the Collins et al. (1999) study, there was a confound with learners' level, with more academically gifted students attending the more intensive course. Finally, while the distribution of course hours was manipulated, there was no attempt to systematically distribute the presentation, practice or review of particular language items or skills. However, the question remains as to whether shorter lags are more effective for certain aspects of language learning.

2.4.3 L2 Oral Fluency and Task Repetition

Yet another area of L2 learning in which there appears to be less of a benefit for distributed practice is oral task repetition (Bui et al.,2019; Kobayashi, 2022; Suzuki, 2021; Suzuki et al.,2022; and Suzuki & Hanzawa, 2022). Task repetition, on its own, has been shown

to improve fluency (De Jong & Perfetti, 2011; De Jong & Tillman, 2018; Lambert et al. 2017; Mackey et al., 2007) but until a recent glut of studies, the optimal length of time between task repetitions was not well researched. Bui et al. (2019) found that massed task repetition of a picture description task benefitted fluency, while longer lags of a week benefited complexity. Suzuki (2021) and Suzuki et al. (2022) found that blocked practice, in which the same narrative task was repeated three times immediately, followed by two more tasks repeated three times over the next two days, compared to an interleaved schedule, in which on each of the three days the three different narrative tasks were performed, improved some measures of fluency and the reuse of lexical constructions. Kobayashi (2022) found only one difference between test performance on an oral narrative task administered on a 7-day RI delayed test after either massed or 7-day ISI. Lexical variety was improved in the spaced condition but not grammatical complexity and accuracy nor fluency. Suzuki and Hanzawa (2022) found that an oral picture narration task repeated six times in a massed condition compared to either a short lag (45-min ISI after three repetitions) or a long lag (7-day ISI after three repetitions) improved fluency measures during training and in an immediate posttest but not in a 7-day RI delayed task. Taken together, these studies suggest that for more procedural, fluency-based tasks, shorter lags and even massed schedules may be better. Bui et al. (2019) drew on Levelt's (1989) speech production model to suggest that massing practice allowed for less attention to be spent on conceptualising the message and more on formulating, articulating and self-monitoring. However, as Rogers (2022) points out, very few of these studies had delayed tests, and those that had one did not include an ISI/RI ratio recommended by the distributed practice literature of between 10-30% (Rohrer & Pashler, 2007, Cepeda et al., 2008). Indeed, one of the consistent findings in the spacing literature is that massed schedules tend to produce higher results on immediate posttests

(Delaney et al., 2010). Further studies may therefore be needed to rule out the possibility that measures of fluency benefit from distributed practice when lags conform to the ISI/RI ratios in the 10-30% range.

2.4.4 L2 Grammar Learning

L2 grammar and the distributed practice effect is less well-researched than vocabulary (at least vocabulary learned in paired-associate experiments) and as yet no consensus has been reached. The first study to be published investigating the distributed practice effect and L2 grammar was carried out by Bird (2010). Bird conducted a classroom-based study with 38 Malay university students enrolled on an English language course learning English verb tenses (past simple vs. present perfect and present perfect vs. past perfect) under intentional conditions over five study sessions. Participants were divided into two groups. Within-subjects factors included ISI (3-day and 14-day), task (past simple vs. present perfect and present perfect vs. past perfect) and RI (7-day and 60-days). One group studied past simple vs. present perfect with a 3-day ISI and present perfect vs. past perfect with a 14-day ISI, while the task-ISI relationship was counterbalanced in the other group. Both groups were given a pre-test and then two surprise posttests at a 7-day RI (ISI/RI ratio of 42% for 3-day ISI and 200% for 14-day ISI) and a 60-day RI (ISI/RI ratio of 5% for 3-day ISI and 23% for 14-day ISI). Only the 14-day ISI, 60-day RI condition at 23% fell within the 10-30% range of optimal ratios suggested by Rohrer and Pashler (2007), although the 3-day ISI, 7-day RI condition at 42% was found to be the similar to the optimal spacing for recall tasks in the experimental data from Cepeda et al. (2008) but outside the 24% optimal ratio (1-day ISI) for recognition tasks. Interestingly, both recognition (deciding whether a sentence was correct) and recall tasks (correcting incorrect sentences) were involved in Bird's study design. Both study and test material involved identifying correct (5) or incorrect

(15) sentences on a worksheet and then correcting the ungrammatical sentences. Feedback was given in the study phase. Results showed that for the longer 60-day RI, scores for the optimal ISI/RI ratio of 14-day ISI (23%) were significantly better than the sub-optimal 3-day ISI (5%) schedule. For the shorter 7-day RI, neither ISI schedule (both sub-optimal according to Rohrer and Pashler, 2007) significantly outperformed the other. Bird concluded that L2 grammar benefits from distributed practice in a similar way, including the ISI/RI ratio, to that found in other domains.

Rather than comparing longer and shorter lags and the ISI/RI ratio, Miles (2014) compared massed with distributed instruction of L2 grammar. Thirty-two Korean university students were split into two groups and given instruction and practice on the word order of English frequency adverbs. In the massed group, participants received the 65 minutes of instruction and practice in one session, while in the distributed group, the same 65 minutes were spread over three sessions with an expanding schedule (7-day ISI between session 1 and 2, and 28-day ISI between session 2 and 3. A pre-test, an immediate posttest and a delayed posttest with an RI of 35 days (ISI/RI ratio of <1% for the massed group for the delayed posttest; ISI/RI ratio for the distributed group of 20%-80% for the delayed posttest) were administered. Tests consisted of an error correction task in which participants were presented with incorrect sentences and asked to correct them, and a more productive translation task, in which Korean sentences needed to be translated into Korean. Results showed there was a significant advantage for the distributed group over the massed group on the error correction task but not on the more productive translation task.

In another classroom-based study investigating longer and shorter lags, Rogers (2015) investigated whether there was a distributed practice effect for adults exposed to L2 syntax (five

different cleft sentence patterns) over five sessions under incidental learning conditions. Participants were presented with cleft sentences one at a time on a screen and asked comprehension questions to focus attention on the meaning rather than the form. Participants were divided into two groups (2.3-day ISI and 7-day ISI). Both groups were administered a pre-test, an immediate posttest, and a delayed posttest with an RI of 42 days (ISI/RI ratio of 5% for the 2.3-day ISI, 42-day RI; 17% for the 7-day ISI, 42-day RI), comprising grammaticality judgment tests. He found that the 7-day ISI group outperformed the 2.3-day ISI group on the 42-day RI but on the immediate posttest there was no significant difference. This study provides evidence of a distributed practice effect for L2 grammar under incidental, albeit far from naturalistic, learning conditions. While the learning conditions were incidental in that the learners' attention was directed away from the form of the sentences and towards meaning via comprehension questions, it would be surprising if the learners did not focus on form as well, given that the sentences were presented one at a time on the whiteboard in a language class. If this were the case, it is possible that it is this explicit focus on form that benefitted from distributed practice. It remains unclear whether L2 grammar learned under more naturalistic incidental conditions benefits from distributed practice. Taken together, these three studies point towards L2 grammar benefitting from the distributed practice effect. However, several other studies found the opposite or no advantage for more distributed practice.

Other studies into L2 grammar learning have found no advantage for longer lags or an actual advantage for shorter lags (Suzuki & DeKeyser, 2017a; Suzuki, 2017; Kasprowicz et al., 2019). Suzuki and DeKeyser (2017a) investigated the learning of Japanese morpho-syntax (to indicate an action similar to the present continuous tense in English) over two sessions with ISIs of either 1 day or 7 days. The two learning sessions lasted between 45-50 minutes and included

both explicit instruction and communicative practice. Two posttests, in the form of two productive tasks (rule application and picture narration task), were given to each participant after 7 days (ISI/RI ratio of 14% for the 1-day ISI, and 100% for the 7-day ISI) and after 28 days (ISI/RI ratio of 3% for the 1-day ISI, and 25% for the 7-day ISI). Suzuki and DeKeyser found no difference between the longer and shorter lag groups and indeed found that the massed learning condition resulted in quicker reaction times in the picture sentence completion task. They concluded that the lack of benefit for the longer lag at ISI/RI ratios suggested by Rohrer and Pashler (2007) for the accuracy measurements and a benefit for the shorter ISI on reaction time measurements was due to the complexity of the item and task. One problematic issue with Suzuki and DeKeyser's methodological design was that learners had prior experience of learning the target structure, some as long as a year or more before. This may have confounded the results by in effect creating a contracting schedule with huge ISIs.

Partially in order to rectify the methodological flaw in Suzuki and DeKeyser (2017a), Suzuki (2017) replicated the study with an artificial language. Sixty participants studied simple and complex morphosyntax over four training blocks with ISIs of either 3.3 days or 7 days. They were tested after an RI of 7 days (ISI/RI ratio of 47.1% for the 3.3-day ISI, and 100% for the 7-day ISI) and 28 days (ISI/RI ratio of 11.8% for the 3.3-day ISI, and 25% for the 7-day ISI) on vocabulary, rule application and sentence picture narration task. Suzuki found that the shorter lag (3.3-day ISI) outperformed the longer lag (7-day) group on accuracy but not on reaction time measurements. However, another methodological issue, which the author admitted, arose with this study. By including two delayed tests, Suzuki was in fact giving an extra chance to practice and this may well have affected the ISI/RI ratios to the point where both were within a similar range of optimal spacing. A re-analysis to include the potentially confounding posttest as

another study session showed that the shorter, now 4.25-day ISI and 21-day RI with an ISI/RI ratio of 20%, were not far off the optimal ratio of 17% for a 35-day RI from Cepeda et al. (2008).

Another classroom-based study by Kasprowicz et al. (2019) looked at the learning of French verb inflections by 113 beginner-level eight-year-old children. Studying either six 30-minute sessions with a 3.5-day ISI, or three 60-minute sessions with a 7-day ISI group on a 42-day RI delayed test (ISI/RI ratio of 8.3% ISI/RI for the 3.5-day ISI, and 16.7% for the 7-day ISI) that involved sentence-picture matching task and an acceptability judgement task, similar to Bird (2010). Results showed that there was no significant difference between the longer and shorter lag groups for either test task. This may be partly to do with the low level of learning by both groups, but also the potential confound of a different number of sessions for the two groups (3x60-minute vs. 6x30-minute sessions).

Taken together, the findings from these studies into the distributed practice effect and L2 grammar are somewhat puzzling in their lack of consensus. However, a closer inspection of the numerous methodological differences between the studies may help our understanding. Some of the different contextual and methodological factors include: the age of the learners (children or adults); level of the learners (beginner to intermediate); the number of sessions (2-5); the amount of time included in the exposure sessions (45 minutes to 360 minutes); spacing schedules (equal vs. expanding schedules); the type of grammar (syntax or morphosyntax); complexity of the language item (simple or complex); different length of lags (massed, 1-day, 3.3-day, 3.5 day, 7-day); retention interval to the delayed test (7-day, 28-day, 35-day, 42 day); the ratio between ISI and RI (less than 1% up to 200%); exposure condition (incidental and intentional); exposure tasks (instruction, communicative practice, both receptive and productive); presence of error feedback (none or some); the type of test task (from more receptive to more productive:

acceptability judgement tasks and correction tasks, picture matching tasks, translation tasks, picture narration tasks); the test measurement (accuracy and reaction time).

2.4.5 Summary of Distributed Practice Studies in SLA

In sum, the benefit of distributed practice for L2 grammar, L2 vocabulary under more naturalistic learning conditions and fluency measures of L2 oral production is less clear cut than for L2 vocabulary learned under paired-associate rote-learning conditions. One of the main challenges facing the small SLA distributed practice research community is to consider why there have been varying results. It is important to determine which of the factors above (or others not mentioned) affect, either on their own or more likely as interactions with each other, the optimal lag for a variety of L2 aspects of language and tasks. In the following section, several of the factors listed above, which can be considered as potential candidates for narrowing the optimal lag and the optimal ISI/RI ratio, will be examined in greater detail.

2.5 The Optimal Spacing of Distributed L2 Grammar Practice

2.5.1 ISI/RI Ratio Rule of Thumb

One possible explanation for why L2 grammar learning distributed practice studies have produced varying results is that the optimal lag for any given RI may be different for L2 grammar learning than for paired-associate vocabulary learning. If this were the case, then studies that have designed experiments with groups with an optimal ISI/RI of 10-30% to fit the range advocated by Rohrer and Pashler (2007) and other non-optimal groups with ISI/RI ratios outside of the 10-30% range (e.g., 8.3% for Kasprowicz et al., 2019) may find that results do not demonstrate the possible difference in test results as the “non-optimal” group may in fact be closer to optimal than originally conceived.

There is evidence to suggest that several factors may influence both the optimal ISI and the ISI/RI relationship for L2 learning, and this may lead to different optimal ratios to those found by, for example, Cepeda et al. (2008) (see Table 1 for a summary). Some of the factors may be related to the ease with which an item is originally encoded in memory (item difficulty, declarative tasks, vocabulary, explicit instruction); some of the factors may be related to how well an item may be maintained (and possibly strengthened) in memory (declarative memory, declarative tasks); and other factors may be related to how well an item is retrieved from memory (free recall test/productive tests). In the following section I shall discuss how these factors may influence optimal spacing.

Table 1. Factors that may influence the optimal spacing and the optimal ratio of ISI to RI

Wider optimal spacing	Narrower optimal spacing	Reference
Simple items	Complex items	Donovan and Radosevich (1999) Janiszewski et al. (2003)
Declarative task	Procedural task	Kim et al. (2013)
Vocabulary (arbitrary)	Grammar (rule-governed)	Ullman & Lovelett (2018)
Explicitly inputted	Implicitly inputted	Janiszewski et al. (2003)
Productive test	Receptive test	Cepeda et al. (2008)
Free recall	Cued recall	Cepeda et al. (2008)
Good declarative Memory	Bad declarative memory	Ullman & Lovelett (2018)

2.5.2 Declarative vs. Procedural Tasks

Declarative memory is the long-term memory system responsible for storing episodic and semantic knowledge. It is a fast-learning, flexible system that can learn both explicitly and

implicitly, although it is thought to be the only system responsible for explicit knowledge (see Ullman, 2016 for an overview). In the field of SLA, it has not only been linked to the acquisition of vocabulary but also, at least partly due to its relative speed compared to procedural memory, the initial stages of grammar learning (Hamrick 2015; Morgan-Short et al., 2014; Ullman, 2004). Procedural memory, on the other hand, is the long-term memory system that is responsible for a wide-range of cognitive and motor-skills. It is slower yet more robust than declarative memory and is always implicit (Ullman, 2004). It is thought to be involved in pattern recognition and habit formation in general. In language acquisition it has been hypothesised to be involved in grammar learning, including non-idiosyncratic aspects of vocabulary learning (Ullman, 2004). Studies have also suggested that the slower-learning procedural memory takes over from declarative memory at later stages of the acquisition process (Hamrick, 2015; Morgan-Short et al., 2014). There is also evidence that procedural memory and declarative memory systems are somewhat redundant in that items can be learned using either and often both systems simultaneously and that learning conditions can affect which system takes a lead role (Ullman & Lovelett, 2018). For example, explicit learning conditions may encourage reliance on the declarative memory system while more implicit conditions may force the use of procedural memory systems (Ullman, 2016; Ullman & Lovelett, 2018).

As Ullman and Lovelett (2018) pointed out, very few distributed practice studies make an explicit link to declarative and procedural memory systems. However, many studies into the distributed practice effect have been carried out with the learning of idiosyncratic information, that is, the memorising and retrieval of declarative knowledge (e.g., Bahrick et al., 1993; Cepeda et al., 2008). With regards to procedural tasks, there is also evidence for a distributed practice effect. Non-linguistic tasks that are thought to rely on procedural memory have been found to

benefit from distributed practice, including serial reaction time tasks (Kwon et al., 2015), fine motor-skill learning (Mackay et al., 2002; Moulton et al., 2006), music performance (Simmons, 2012) and completing complex mathematical problems (Rohrer and Taylor, 2006). Kwon et al. (2015) compared three sessions of a serial reaction time test with gaps of either 10 minutes or 12 hours (including sleep). The posttest was administered immediately after the final block of training. Results showed that distributed practice improved reaction times during training and at the posttest compared to the much shorter lag. However, it is important to note that there was no delayed posttest with a retention interval in the range laid out by Rohrer and Pashler (2007) or even Kornmeier et al. (2019). Mackay et al. (2002), investigating surgical skill performance, found that distributed practice (20-minute ISI), compared to massed practice, influenced gains made during practice and not just on delayed test (5-minute ISI, ISI/RI ratio of 400%). Again, however, the retention interval to the posttest of only 5 minutes gave an ISI/RI ratio significantly over the optimal range of 10-30% laid out for more declarative tasks (Rohrer & Pashler, 2007). An early meta-analysis by Lee and Genovese (1988) on motor-skill learning tasks suggested that relative gains made during training in a distributed practice condition compared to a massed condition remained, albeit diminished, after a retention interval. One example of distributed practice giving benefits over a more educationally relevant retention interval comes from Moulton et al. (2006), who investigated the development of surgical skills. They had two groups: massed in one day and distributed over seven days. They included an immediate posttest and a delayed test (M=28.5 days, ISI/RI ratio of 25% for the distributed group). Results showed that, contrary to Mackey et al., (2002), there was no difference in performance on the immediate posttests, but on the delayed posttests, the distributed group outperformed the massed group on nearly all measures. In a meta-analysis of seven studies investigating distributed practice for

simulator-based training for surgical skills, Fahl et al. (2023) concluded that distributed practice was more effective than massed practice, but 1-day ISIs were generally better than 7-day ISIs. They suggested that either reactive inhibition (fatigue, boredom) or consolidation factors may be responsible for the distributed advantage. Evidence from sleep-consolidation studies has demonstrated off-line learning of skill acquisition (e.g., Robertson et al., 2004), specifically due to reactivation during sleep (Schönauer et al., 2014). The potential benefits of sleep (but not other potential mechanisms) are supported by another surgical skills study by Bjerrum et al. (2016), who included a delayed posttest of 28-day RI) but found no difference in performance between a one-day ISI and a 7-day ISI.

Despite the findings from complex skill learning and motor skill learning studies, a couple of studies from a skill acquisition perspective have cast doubt over the benefit of distributed practice for procedural memory systems (Kim et al., 2013; Paik & Ritter, 2016). Similar to other skill acquisition theories (e.g., Anderson, 1982; DeKeyser, 2020; Fitts, 1964), Kim et al. (2013) outlined their theory in which skills moves from a stage in which there is a reliance on declarative memory through to a declarative-procedural stage, ending with a procedural memory stage. They posit that declarative memory is forgotten at a faster rate than procedural memory, and therefore the effect of distributed practice will differ depending on the stage of acquisition. They suggested that distributed practice is more beneficial during the first, declarative stage, while massed practice might be beneficial to move from the second stage to the third, that is, for proceduralisation. Paik and Ritter (2016), in a follow-up empirical study, compared the distributed practice on declarative (English-Japanese word learning task), procedural (Tower of Hanoi task) and perceptual-motor tasks (an inverted pendulum task). They concluded that distributed practice only benefited declarative memory and not procedural

memory, as the forgetting rate was much lower with procedural memory. They explained the distributed practice effects found in procedural tasks such as solving mathematical problems (Rohrer & Taylor, 2006) and perceptual motor tasks, such as Moulton et al. (2006), which investigated the development of surgical skills, as only the declarative aspects of the complex skill benefitting while the procedural aspects did not. The acquisition of L2 grammar and vocabulary under more naturalistic settings involves both declarative and procedural knowledge. Applying the theory from Kim et al. (2013) to L2 learning, it is possible that the more declarative aspects of the language (e.g., vocabulary learned by rote, grammar rules that are explicitly taught) benefit more from distributed practice than more procedural aspects (incidentally learned grammar and vocabulary, language production).

An alternative interpretation of the varying findings is SLA distributed practice studies is that there is a qualitative difference in the distributed practice effects regarding declarative and procedural tasks. One interesting difference between distributing practice on procedural and declarative tasks is on the effect of the training phase. Studies from motor-skill learning tend to show improvement during learning (e.g., Mackey et al., 2002), while with more declarative tasks, learning is often hindered by spacing and it is only on the delayed tests that the benefit of spacing shows (Cepeda et al. 2006). One possible explanation for this is that different underlying mechanisms are at work with each. For procedural tasks, offline consolidation particularly after periods of sleep (i.e., around the 1-day ISI) may be the main factor (Brown & Robertson, 2007; although see Rakowska et al., 2021 for evidence of offline gains for a serial reaction time task over six weeks). For declarative tasks, it is possible that a combination of reminding, contextual variability and consolidation play a role.

2.5.3 Complexity

In a meta-analysis of distributed practice studies by Donovan and Radosevich (1999) more complex items were found to have a weaker spacing effect than simple items. However, as Rogers (2017) points out, this should be taken with caution, as the categories Donovan and Radosevich used to define complexity were different from those used in SLA, including tasks that include physical but not mental complexity. Janiszewski et al. (2003) in their own meta-analysis found that there was a larger spacing effect for semantically but not structurally complex stimuli. This leaves open the possibility that the distributed practice effect plays less of a role in L1 and L2 learning than in rote or even conceptual learning. Indeed, in Suzuki and DeKeyser's (2017a) study into the distributed practice effect and L2 Japanese morphosyntax, they found that there was no difference between accuracy scores in the massed and distributed groups. One possible explanation they gave for these findings was that the oral productive test used in this study was much more complex than the more receptive tasks used in previous L2 grammar studies (Bird, 2010; Rogers, 2015).

2.5.4 Productive vs. Receptive Tests

The fact-learning distributed practice study of Cepeda et al. (2008) used a free recall and a multiple-choice test, which the authors termed recall and recognition respectively but which from an SLA perspective could possibly be considered productive and receptive tests. As can be seen from Figure 3., for the comparatively shorter RI of 7 days, the optimal ISI ratio differed greatly between free recall and recognition. For free recall the optimal ISI/RI ratio was 43% or 3 days; for recognition, the optimal spacing was 1.6 days or a ratio of 24%. As the RIs got longer, the optimal ratios between recall and recognition tests converged. One explanation for this finding comes from contextual variability theory. It has been posited to work at shorter ISIs and

RIs but less at longer ones as contextual factors found in the delayed posttest and in the training phase are more likely to overlap after shorter retention intervals than after long retention intervals (Glenberg, 1979), and it is possible that this mechanism interacts with whether the delayed test is recall or recognition. Indeed, Glenberg (1979) suggests that contextual cues do not help as much on recognition tests as these types of tests already provide contextual cues, whereas recall tests may rely more on the contextual cues encoded with the memory trace. Therefore, at shorter retention intervals, when contextual variability plays more of a role, the difference between recall and recognition is large, but as retention intervals increase and the influence of contextual variability wanes, so the difference between recall and recognition contextual cues are more likely to overlap. There is an alternative explanation from a reminding account perspective. As mentioned earlier, Suzuki and DeKeyser (2017a) claimed that their productive test was more complex and therefore benefited less from the distributed practice effect. This is an area that requires further research.

2.5.5 Intentional vs. Incidental Learning Conditions

A number of experiments have found a spacing effect for incidental learning conditions (Challis, 1993; Glenberg & Lehman, 1980; Greene, 1990; Greene & Stillwell, 1995; Jensen and Freund, 1981; Shaughnessy, 1976; Verkoeijen et al.; 2005), although it took some time before a consensus could be reached regarding several attenuating and potentially confounding factors, including the effect of free recall compared to cued recall tests, the amount of semantic processing and the length of ISI. Greene (1989) found that when tested using free recall tasks, a spacing effect emerged for both intentional and incidental learning conditions, but for a cued recall task, only the intentional condition showed signs of a spacing effect. In a follow-up study, Greene (1990) found a spacing effect on three tests of implicit memory: spelling of homophonic

words, word-fragment completion and perceptual identification. Challis (1993) proposed and subsequently showed that the level of semantic priming affected cued recall tasks. That is, incidental conditions also demonstrate a spacing effect when there is a semantic priming but not non-semantic priming. Toppino and Bloom (2002) repeated Greene's (1989) earlier study, modifying to remove the possibility of a recency effect and found no spacing effect for incidental conditions in a free recall task. However, when they shortened the length of the ISI, the spacing effect re-appeared. The meta-analysis by Janiszewski et al. (2003) showed a stronger effect size for intentional compared to incidental conditions. However, this may be the result of studies that did not control for different ISIs, thus finding a smaller effect size for incidental conditions when tested at the same RI.

In a subsequent study, highly relevant for study 2 of this thesis, Verkoeijen et al. (2005) presented 80 target nouns within a longer list of 160 words one at a time at 10-second intervals. An intentional group were asked to remember the words for a subsequent test. The incidental group were asked to find the rule that determined the order the words were presented. The target nouns were presented two times and spaced at lags of 0, 21, 52.5, 84, 147 or 210 seconds. An immediate posttest was administered in which participants had to recall as many of the words as they could. Results demonstrated an inverted u-shape function of learning. They also showed that intentional learning conditions required a wider optimal lag (147 seconds) than incidental conditions (52.5 seconds), but that at their respective optimal lags, there was a larger distributed practice effect for intentional conditions (57% correctly recalled) compared to the incidental condition (30%). They posited that from a study-phase retrieval account, the deeper processing involved in intentional learning leads to the memory trace lasting longer and therefore strengthening more when it is retrieved at its optimal ISI, that is, close to failure. However, the

lack of a delayed posttest raises questions of whether this applies at longer retention intervals. It is also unclear whether the shorter optimal lag and lower magnitude of distributed practice effect would be the same for more procedural knowledge, for example L2 grammar.

To summarise this series of studies, it appears there is a distributed practice effect for incidental as well as intentional learning conditions but that optimal spacing is narrower and the effect size possibly smaller. However, before generalising to L2 acquisition, it bears reminding that as with most other distributed practice research, the vast majority of these studies involved very short ISIs and RIs in seconds and minutes rather than days and involved declarative knowledge recall tasks rather than more procedural learning.

As has been discussed in previous sections (2.4.1.2 and 2.4.4), several studies have investigated the distributed practice effect under incidental learning conditions in L2 learning, including contextual vocabulary learning (Elgort & Warren, 2014; Macis et al., 2021; Serrano & Huang, 2018; Webb & Chang, 2015) and L2 grammar learning (Rogers, 2015). Of these, only Macis et al. (2021) included both deliberate and incidental learning conditions. As mentioned earlier, they found that distributing the L2 collocations had a large effect on improvements in the deliberate learning condition and a small effect on improvements in the incidental learning condition. Massing the L2 collocations had a medium effect on improvements in the incidental learning condition. These results suggest that a massed conditions, or at least shorter lags, may be more effective under incidental learning of L2 collocations. However, this study did not look at the abstraction and transfer of rules that can be found in L2 grammar.

2.5.6 Abstraction and Transfer

Abstracting rules and patterns and then transferring them to new learning situations lies at the heart of much human learning, from playing computer games to grammar learning. The question of whether spacing out the presentation and practice of to-be-learned items in incidental learning conditions increases the likelihood that generalised rules can be abstracted and then transferred to other learning situations is of considerable interest. Research from various domains suggests that distributing the presentation and practice of more complex tasks improves the transfer of learned rules and procedures to different learning contexts and tasks (Hagman, 1980; Kornell & Bjork, 2008; Moulton et al., 2006; Rohrer & Taylor, 2006; Vlach et al., 2012; Vlach et al., 2008; see Carpenter et al., 2012, for a review). A number of these studies involved intentional learning conditions (e.g., Hagman, 1980; Moulton et al., 2006; Rohrer & Taylor, 2006) in which participants were instructed to learn particular rules and then transfer that knowledge to new examples. Other studies involved abstracting commonalities from exemplars but under intentional learning conditions, that is, having been instructed to try and learn them (e.g., Kornell & Bjork, 2008). Only a few studies have attempted to discover whether the abstraction and transfer of rules under incidental learning conditions benefits from distributed practice (Ambridge et al., 2006; Rogers, 2015; Vlach et al., 2008).

Vlach et al. (2008) used a category-induction word-learning task in which made-up objects that differed in several aspects but kept the same shape were presented to three-year-old children in either massed or spaced conditions. Those in the spaced condition were better able to abstract the core meaning and identify a new version of that object on a multiple-choice posttest. A memory task in which the children were presented with the same object and just had to remember also found that spaced practice was better than massed but produced significantly

better results than the category-induction task. Vlach et al. (2012) proposed a forgetting-as-abstraction theory to explain why distributing the presentation of exemplars promotes abstraction of rules. Upon encountering a second exemplar after a period of time, generalised aspects of the to-be-learned items are strengthened, while non-generalised aspects continue to be forgotten. Therefore, over time, rules are abstracted.

Another study by Ambridge et al. (2006) investigated whether distributing practice of rules under incidental learning conditions better allowed them to be abstracted and then transferred to novel test items. In two experiments, children (M = 5.3 years for experiment 1, M = 4.5 years for experiment 2) were given ten exposures of the past tense object-cleft sentence construction (It was the [OBJECT] that the [SUBJECT][VERB]ed) using glove puppets. Learning schedules were either massed (all ten exposures in one day), distributed (one per day for ten days), or distributed pairs (two per day for five days). The immediate posttest consisted of an elicited production task using verbs that had not appeared in the exposure phase. Results showed that children in the distributed and distributed pairs schedules outperformed those in the massed schedule. It should be noted that the posttest was administered immediately after the last exposure phase, so the results demonstrating an advantage for distributed practice perhaps mirror studies into more procedural tasks that find an advantage during training rather than on a delayed posttest (e.g., Mackey et al., 2002). It is therefore also unclear whether the distributed advantage would hold after a delayed posttest with an RI in the ISI/RI range suggested by Rohrer and Pashler (2007).

All of the L2 grammar distributed practice studies mentioned above (Bird, 2010; Miles, 2014; Rogers, 2015; Suzuki & DeKeyser, 2017a) involved applying the rules that had been given to them to new exemplars in the testing phases, albeit only Rogers (2015) involved incidental

learning conditions in which the learners had to abstract the rules for themselves. However, no study, to our knowledge, has contrasted items previously trained (i.e., a test of memory) and generalisation of rules to new exemplars as within-subject variables to test optimal ISI/RI ratios.

2.5.7 Summary of Factors that may Influence the Optimal ISI/RI Ratio

To summarise this section, there are a number of factors that may influence the optimal ISI/RI ratio, resulting in a difference from the findings of Cepeda et al. (2008) and the 10-30% ISI/RI rule of thumb by Rohrer and Pashler (2007), which was, in turn, based on the data from Cepeda et al. (2008). Determining what some of those factors that influence the optimal spacing may help our understanding of the underlying mechanisms of the distributed practice effect and help make sense of the mixed findings in L2 distributed practice research. In study 2 (chapter 5), I investigated several of these factors (intentional vs. incidental learning conditions; items that appeared in the exposure phase vs. new items that require a generalisation of rules; and declarative memory).

2.6 Individual Differences and the Distributed Practice Effect

2.6.1 Bringing Individual Differences into our Understanding of Distributed Practice

To what extent do individual differences modify the effect of distributed practice? The Multiscale Context Model (MCM) model of Lindsey et al. (2014) has a parameter for individual abilities, recognising that the optimal spacing for individual learners depends to some degree on cognitive differences. However, they do not stipulate which cognitive factors these are, instead allowing for a general ability score. More recently, there has been a call for bringing individual differences into our understanding of the mechanisms of the distributed practice effect (Knabe & Vlach, 2020). In a review of the somewhat scant literature regarding early childhood studies into

the distributed practice effect and the lack of consensus in their results, Knabe and Vlach (2020), call for a more nuanced, individual difference account of the phenomenon. Drawing on studies that show how age-related differences in attention (Slone & Sandhofer, 2017), memory (Vlach & Johnson, 2013) and metamemory (Vlach et al., 2019) may affect the desirable difficulty of a task and therefore the efficacy of spaced or massed practice, Knabe and Vlach (2020) conclude that the next generation of theories into the distributed practice effect need to take into account individual differences. In the following sections, I review the research into individual differences and the distributed practice effect.

2.6.2 Distributed Practice and Cognitive Deficiencies

A search of the literature reveals only a few studies into cognitive differences and spacing, some of which investigated children and adults with cognitive deficiencies. Madsen (1963) found that children with low IQ benefitted more from distributed presentation of paired associates than those with higher IQ, although this finding was not replicated in a later study by Sperber (1974). In a study into dementia, spaced practice has been found to be beneficial for patients (Camp et al., 1996), particularly the use of spaced retrieval techniques, which involves shortening the next ISI if a patient struggles to remember the item being tested or lengthening the next ISI when the patient can recall the item tested (Hawley et al., 2008). Riches et al. (2005) found that spacing the presentation helped the learning of novel verbs for 5-year-old children with specific language impairment, which according to Ullman and Pierpont's (2005) procedural deficit hypothesis is due to damage to the procedural memory system, with working memory also damaged (see also Lum et al., 2012; but see also Goffman & Gerken, 2020 for an alternative explanation). Gettinger et al. (1982) showed that learning disabled children's spelling benefited from spaced practice. These studies suggest that it is not as simple a case as those with better

memories benefit more from spacing and those with worse memories benefit from massed. Study-phase retrieval and the desirable difficulty aspect of the reminding account (Benjamin & Tullis, 2010) suggests that the closer the memory trace is to failure upon retrieval, the more the memory trace will be strengthened. Thus, it is possible that what these studies show is that while distributed practice is beneficial for everyone, the optimal lag may be narrower for those with cognitive and learning difficulties. The findings in Madsen (1963), that those with low IQ benefitted more from distributed practice than those with higher IQs may therefore be explained by the chosen ISI in the experiment being closer to optimal for the lower IQ children.

2.6.3 Individual Differences that Favour Massed or Distributed Practice

Three other studies found evidence for strengths in certain cognitive areas favouring either massed or distributed practice. Mumford et al. (1994) found that those with high visual-spatial ability benefited from spaced practice of a complex cognitive skill, but they were not as successful as those with low visual-special ability under massed practice conditions. In contrast, those with low perceptual speed did not do well under massed conditions. Verkoeijen and Bouwmeester (2008), using latent class regressions on datasets of free recall list learning, found that the benefits of spacing depended not only on ability of the learners but on the presentation speed. Higher-level learners benefited more from spacing when the presentation rate was quick, but lower-level learners benefited more from spacing when it was slow. Elgort and Warren (2014), in their study into incidental L2 vocabulary learning while reading, mentioned in the previous section, found that lower proficiency participants were only able to infer the meaning of the pseudo-words when they were encountered within the same chapter. This provides some support for a desirable difficulty view of the underlying mechanisms of the distributed practice

effect, as the cross-chapter spaced condition may have presented an undesirable level of difficulty for the lower proficiency participants.

2.6.4 Working Memory and Language Analytic Ability

Working memory and language analytic ability have both received some attention with regards to distributed practice. Working memory is the ability to not only hold but also simultaneously process items in short-term memory and measured by complex span tasks. Language analytic ability combines both grammatical sensitivity and inductive learning ability (Shintani & Ellis, 2015) and is often measured by the Llama_F test (Meara, 2005). Suzuki and DeKeyser (2017b) investigated whether working memory, and language analytical ability correlated with longer (7-day) and shorter (1-day) lag schedules when learning Japanese. They found that working memory capacity correlated with the shorter lag schedules while language analytical ability correlated with the longer lag. Working memory capacity has been linked to the noticing of grammatical regularities in language and on-line language processing (Coughlin & Tremblay, 2013). An interpretation of the results of both Suzuki and DeKeyser (2017b) and Verkoeijen and Bouwmeester (2008) is that when to-be-learned items are presented rapidly one after another under massed conditions, participants must be able to hold and process those items quickly and efficiently in their working memory, and therefore those with better working memory capacity will be more likely to succeed under massed conditions. Supporting evidence comes from a study by Chen et al. (2018) into spacing and working memory of primary school students' learning of mathematics problems. They found that massing compared to spacing (1-day lag) depleted working memory resources, and thus reduced the amount of learning. Interestingly, these findings and explanations run contrary to the deficient processing hypothesis (Hintzman, 1974), which predicts that those with better working memory capacity will lose the

benefit of even small ISIs during massed conditions as the to-be-learned items remain in working memory longer. Another L2 grammar study by Kasprowicz et al. (2019) investigated whether language analytic ability interacted with two different lags (ISIs of 3.5 days and 7 days) when learning French verb morphology. They found that while language analytic ability influenced gains in both groups, there was no significant interaction between lag and group.

These studies suggest that distributing practice is not necessarily optimal for everyone and that other factors may interact with individual differences. However, through studying interactions between individual differences and distributed practice schedules, a better understanding of the underlying mechanisms of the distributed practice effect may be gained.

2.6.5 Declarative Memory and Procedural Memory

To the best of my knowledge, and perhaps somewhat surprisingly, only a couple of studies have so far been carried out regarding the distributed practice effect and declarative and procedural memory systems as individual differences, even though for declarative memory, at least, predictions based on underlying mechanisms of the distributed practice effect can be made. For declarative memory, study-phase retrieval accounts, and its descendent, the reminding account, predict that the optimal spacing (ISI) of paired-associate or other idiosyncratic items should depend partially on the learner's declarative memory ability as the closer an item is to not being successfully retrieved, that is, an optimally desirable level of difficulty, the more the memory trace is strengthened when it is. Therefore, according to this theory, those with better declarative memory should require wider optimal lag than those with worse. Contextual variability theories, on the other hand, do not predict that declarative memory should influence optimal lag.

One study that has investigated the potential interactions between distributed practice and declarative memory is a study by Li (2017). Eighty-six L1 speakers of English were taught and provided practice of L2 Mandarin vocabulary, including their associated tonal pronunciation across five sessions with ISIs of either 1 day or 7 days and RIs of either 1 week (ISI/RI ratio of 14% and 100%) or 4 weeks (ISI/RI ratios of 3.6% and 25%). Declarative memory was assessed via the Continuous Visual Memory Task (Trahan & Larrabee, 1983). Results showed no significant interaction between declarative memory and distributed practice (either ISI or RI). These results did not support a reminding account view of the distributed practice effect..

Perhaps due to the lack of consensus around distributed practice and more procedural tasks such as L2 grammar learning, it is less straightforward to make predictions about how procedural memory as an individual difference might interact with distributed practice than for declarative memory. Despite this, and the fact that language learning (both L1 and L2) differs significantly from other non-declarative tasks such as motor-skill learning, Ullman and Lovelett (2018) postulated that individual differences in procedural memory might interact with distributed practice to affect language learning performance both during training and on a delayed test. However, even if this were the case, it remains unclear whether the interaction would be similar for grammatical and lexical items, and at what stage of the learning process procedural memory system would emerge as the driving force in the acquisition process. Differences in procedural memory strength may impact not only the speed of learning of a more procedural task during the learning phase but also provide additional offline consolidation during the lag periods.

2.6.6 Summary of Distributed Practice and Individual Difference Studies

In summary, no clear picture has yet emerged from the few studies that have been carried out into distributed practice and individual differences. It would appear that distributing practice may be better for those with certain cognitive profiles, while massing practice may be better for others. From a reminding account / desirable difficulty perspective, the ISI should be optimised so that the learner is taxed to the 'optimal' degree, which could mean, for example, that those with worse declarative memory might benefit from shorter ISIs than those with better declarative memory (; Benjamin & Tullis, 2010; Delaney et al., 2010). Individual differences and distributed practice is an area that requires further investigation. In addition to helping computational models such as the MCM by Lindsey et al. (2014) better optimise individual spacing, it could also help both teachers and individual learners design their courses and study plans more effectively**.**

2.7 Summary of Literature Review

In this chapter, I first reviewed studies into incidental and implicit learning in SLA in order to situate the two studies in this thesis within the broader research field in SLA. I defined the distributed practice effect, including the lag effect. Next, I discussed the underlying mechanisms of the distributed practice effect. I then reviewed the burgeoning literature of L2 distributed practice studies with a particular focus on incidental learning conditions. Finally, I reviewed the small literature around the interaction between distributed practice and individual differences. A number of questions have emerged. It is not clear whether incidental learning of L2 grammar and vocabulary under cross-situational learning conditions benefits from distributed practice. Nor is it clear whether learning under cross-situational learning conditions is durable after a retention interval of a day or longer. Another gap in the literature is the need to determine

whether the optimal lag is affected by various factors including intentional and incidental learning conditions, individual memory differences and whether learning is abstracted and transferred to new contexts. Through investigating these questions, a better understanding of the underlying mechanisms of the distributed practice effect may be gained.

The next two chapters report study 1, which investigated the learning of an artificial language under cross-situational learning conditions. Chapter 3 focuses on the order of acquisition, the durability of learning over 24 hours, the interrelatedness of learning and the role that five individual memory difference measures play. Chapter 4 reports on a reanalysis of the data with regard to distributed vs. massed spacing schedules and how the five individual memory differences interact with the different schedules. Both chapters are organised with an introduction, method, results and discussion.

Chapter 3: Study 1. Distinctions in the Acquisition of Vocabulary and Grammar: An Individual Differences Approach

3.1 Introduction

This chapter comprises a published study in the journal, Language Learning (Walker et al., 2020), which investigated the learning of an artificial language made up of nouns, verbs, adjectives, case markers and a verb-final word order under incidental, cross-situational learning conditions. This chapter particularly focuses on the order of acquisition, the durability of learning after 24 hours, and the interaction of five individual memory differences in the learning and retention of this language. This article did not, and therefore this chapter does not, include a description and analysis of how distributed and massed practice schedules compared, nor how the individual memory differences interacted with exposure schedule on learning and retention. That part of study 1 will be included in chapter 4. The chapter is organised exactly as the published article with a literature review, method, results, discussion and conclusion.

3.2 Review of Relevant Literature

A Syrian refugee claims asylum in Sweden; the child of a Chinese economic migrant starts her first day of school in Canada; a British tourist passes through Turkish customs at the beginning of a fortnight's holiday. In each of these examples the learner may know almost nothing of the local, non-native language. The early stages of second language learning under such immersion conditions entail a great deal of ambiguity as learners struggle to make sense of the stream of input they hear by detecting word boundaries, decoding the meanings of words, identifying lexical categories, and understanding the relations between categories defined by the syntax. Even at this early stage, there is individual variation in the ease with which learners pick up the new language (see Dörnyei, 2014 for an overview). How learning is achieved and how individual differences may affect the learning process have been critical questions in the

cognitive sciences (Frost & Monaghan, 2016; Marcus, 1996; McGregor et al., 2005; Siegelman et al., 2017).

Recent research has shown that it is possible for children and adults to learn vocabulary within basic categories of words when they are presented across multiple ambiguous learning situations without any feedback, a process known as cross-situational learning (Yu & Smith, 2007). Smith and Yu (2008) showed that 12- to 14-month-old infants could learn the meanings of novel nouns by keeping track of cross-trial statistics. Scott and Fisher (2012) further demonstrated that 2.5-year-old toddlers could utilise distributional cues to learn novel verbs. With adult participants, Monaghan and Mattock (2012) found that function words could aid the cross-situational learning of nouns and verbs compared to an artificial language where function words did not co-occur with grammatical categories in the language. Then, in a follow-up study, Monaghan et al. (2015) found that nouns and verbs could be learned simultaneously without syntactic cues. The learning mechanisms underlying cross-situational learning are still subject to debate, with some theories proposing an associative, accumulation of statistical probabilities (Yu & Smith, 2007) and others putting forward hypothesis-testing accounts (Medina et al., 2011). A recent study by Khoe et al. (2019) modelling the two approaches suggests that with more ambiguity in the learning environment, learning is more associative. These associative statistical learning mechanisms may paradoxically be domain-general yet also constrained by, and therefore distinctly represented within, different modalities (Frost et al., 2015). While statistical learning has tended to be examined with word learning tasks (e.g., Smith & Yu, 2008), implicit learning has been examined with artificial grammars (e.g., Gómez & Gerken, 1999). Studies such as that of Monaghan et al. (2015) mentioned above and that reported here draw together the two research traditions (see also Monaghan et al., 2019).

These studies looking at the cross-situational learning of nouns and verbs, however, still entail substantial abstraction from the complexity of natural language acquisition. With every new word category or syntactic phrase added, the number of possible referents for any given word increases, making the tracking of statistical probabilities more complex. In a recent study utilizing a novel paradigm, Rebuschat et al. (2021) demonstrated that it was possible for adults to learn a more complex artificial language under cross-situational learning conditions. The artificial language consisted of a verb-final syntax and contained nouns, verbs, adjectives, and case markers denoting the agent and patient of the sentence. Participants saw two dynamic scenes on a computer screen and heard a sentence in the artificial language. Their task was to decide which scene the sentence referred to, and no feedback was given as to whether the participant was right or wrong. Rebuschat et al. (2021) observed that verbs and basic word order were learned first, followed by nouns, then adjectives and finally case markers, which is in line with first language acquisition studies into languages that can omit subjects (e.g., Korean: Choi, & Gopnik, 1995; Mandarin: Tardif, 1996) and adult second language (L2) first exposure studies, which demonstrate the increased salience of sentence-initial and sentence-final positions (Fernald et al., 1992; Shoemaker & Rast, 2013). Studies such as these tentatively suggest that this novel cross-situational learning paradigm may be a useful proxy for the early stages of language learning under immersion settings for L2 adults in future research.

Whereas Rebuschat et al. (2021) demonstrated the viability of a sentence-to-scene cross-situational learning, two key questions about natural language acquisition remain unanswered by their investigation. Firstly, the training and testing used by the authors took place in a single session, and the ability of participants to retain both syntactic and vocabulary knowledge over time is not yet known. If it is possible to learn more complex language through cross-situational

learning conditions, is that learning durable? To our knowledge, only one study has investigated the long-term learning effects of cross-situational word learning. Vlach and Sandhofer (2014) found that noun learning under cross-situational conditions was still robust one week later. In related research, several studies have demonstrated the durability of statistical learning (Durrant et al., 2011; Gómez et al., 2006; Kim et al., 2009); Vuong et al., 2016). Durrant et al. (2011), using Saffran, Aslin, and Newport's (1996) paradigm, showed that statistical learning lasts 24 hours and benefits from sleep consolidation, while Kim et al. (2009) found similar robustness in a visual statistical learning task. To this end, the current study tests learners' knowledge of vocabulary and syntax immediately after training, but also after a 24-hour delay. Second, we note that the relations between learning syntax and vocabulary have been underexplored. It may be that acquisition of vocabulary and syntax are associated with different processes, as reflected by distinct sets of individual differences in learning and memory.

3.2.1 Relations Between Vocabulary and Syntax

The chicken and egg problem of learning syntax and vocabulary has led to proposals either for independence of learning grammar and vocabulary (e.g., Marcus, 1996; Peña et al., 2002), or their inter-relatedness (e.g., Bates & Goodman, 1997; Frost & Monaghan, 2016). In other words, are the referents for vocabulary items (nouns, verbs, adjectives) learned in the same way as grammatical items (word order, case markers) or do they depend on different mechanisms? Many previous studies of artificial language learning have trained participants on vocabulary before testing them on grammatical structure (e.g., Friederici et al., 2002; Morgan-Short et al., 2014; Williams & Kuribara, 2008), and neuropsychology patient studies (Alario & Cohen, 2004), theoretical models (Bock & Levelt, 1994), and memory models (Ullman, 2004) have tended to treat vocabulary and syntax as distinct. An alternative is that grammar and

vocabulary instead depend on a single, domain-general learning mechanism (Bates & MacWhinney, 1987; Frost & Monaghan, 2016; MacDonald et al., 1994; Rumelhart et al., 1986; Seidenberg, 1997). Whether vocabulary and grammar are related to the same, or different, patterns of cognitive abilities, enables a test of whether learning is coherent or fragmented into one or several abilities.

3.2.2 Individual Differences in Cross-Situational Learning

As we suggested above, acquisition of vocabulary and syntax may be differentially sensitive to individual differences in memory. In this paper we will consider a number of types of memory, namely phonological short-term memory (PSTM), working memory capacity, declarative memory, and procedural memory. Phonological short-term memory, the short-term store for auditory information and articulatory rehearsal as measured by simple span tasks, has been implicated in vocabulary acquisition (Baddeley et al., 1988; Gupta, 2003; Martin & Ellis, 2012; Papagno et al., 1991), and grammar abstraction (N. Ellis, 2012; Robinson, 1997; Speidel, 1993; Verhagen & Leseman, 2016). Working memory capacity, defined as the ability to not only hold but also simultaneously process items in short-term memory and measured by complex span tasks, has been linked to the noticing of grammatical regularities in language and on-line language processing (Coughlin & Tremblay, 2013; Mackey et al., 2002; Mackey et al., 2010; Sagarra & Herschensohn, 2010). Sagarra and Herschensohn (2010), for example, found that, in low-level L2 learners, working memory capacity predicted success on grammaticality judgment of gender agreement. A meta-analysis by Linck et al. (2014) found that working memory capacity was a better predictor of a range of L2 performance measures than simple, storage-only memory such as phonological short-term memory. However, a consensus has yet to be reached about the role of working memory capacity under incidental learning conditions, with some

studies reporting a positive relationship between working memory capacity and learning outcomes (Robinson, 2005; Soto & Silvanto, 2014), and others showing no effect of working memory capacity (Grey et al., 2015; Hamrick, 2015; Tagarelli et al., 2015). In a study on implicit and explicit corrective feedback, Li (2013) also found evidence to suggest that L2 proficiency could mediate the impact of working memory capacity, with lower-proficiency learners less reliant on it than higher-proficiency learners.

Declarative memory is the long-term memory system responsible for storing episodic and semantic knowledge. It is a fast-learning, flexible system that can learn both explicitly and implicitly, although it is thought to be the only system responsible for explicit knowledge (see Ullman, 2016, for an overview). In the field of second language acquisition, it has not only been linked to the acquisition of vocabulary, but also, due partly to its relative speed compared to procedural memory, the initial stages of grammar learning (Hamrick, 2015; Morgan-Short et al., 2014; Ullman, 2004). In contrast, procedural memory is a non-declarative long-term memory system. It is slower to learn, yet more robust than declarative memory, and it is always implicit (Ullman, 2004). It is thought to be involved in pattern recognition and habit formation in general. In language acquisition, it has been hypothesised to be involved in grammar learning, including non-idiosyncratic aspects of vocabulary learning (Ullman, 2004). Studies have also suggested that slower-learning, procedural memory takes over from declarative memory at later stages of the acquisition process (Hamrick, 2015; Morgan-Short et al., 2014). There is evidence that procedural and declarative memory systems are somewhat redundant in that items can be learned using either or both systems simultaneously, and that learning conditions can affect which system takes a lead role (Ullman & Lovelett, 2016). For example, explicit learning

conditions may encourage reliance on the declarative memory system while more implicit conditions may force the use of procedural memory (Ullman, 2016; Ullman & Lovelett, 2016).

To our knowledge, only two studies have looked specifically at individual differences in memory in cross-situational learning. Schoetensack (2015) found that neither working memory capacity nor phonological short-term memory predicted cross-situational learning of nouns and verbs under incidental learning conditions, although she reported that in the instructed condition the relationship between phonological short-term memory and learning approached significance. Vlach and DeBrock (2017) found that a paired-associates memory test predicted success on a cross-situational word-learning task in young children. They further demonstrated that visual, auditory, and word-binding declarative memory each play a role.

On the other hand, a number of studies have investigated statistical learning ability, which is the capacity to implicitly keep track of statistical information in the input to acquire linguistic information, as an individual difference in its own right (see Siegelman et al., 2017, for an overview). Statistical learning ability is most commonly measured by means of the auditory speech segmentation task of Saffran et al. (1996) or by means of a visual statistical learning task (Fiser & Aslin, 2002; Kirkham et al., 2002). It has been found to predict success in a number of aspects of language learning relevant to the current study. Firstly, several studies have shown that statistical learning predicts vocabulary development (Shafto et al., 2012; Singh et al., 2012). Shafto et al. (2012) used a visual sequence learning task with 8-month-old infants and found that it predicted success on vocabulary comprehension tests. Meanwhile, Singh et al. (2012) also tested similar-aged infants and found that an auditory word-segmentation task predicted the size of productive vocabulary at 24 months. In addition, statistical learning is linked to sentence processing. For example, Misyak and Christiansen (2012) showed that adults' performance on

auditory statistical learning tasks was the main predictor for L1 sentence comprehension. Finally, statistical learning has also been shown to predict success in the acquisition of syntax. Kidd (2012) found that statistical learning ability, as measured by a visual sequence learning task, predicted the learning of syntax in four and five-year-old children and that this was still the case on a delayed test 24 hours later. In a later study, Kidd and Arciuli (2016) showed that variability in six-to-eight-year-olds' comprehension of aspects of syntax was predicted by their performance on a visual sequence learning task. Interestingly, as in several other studies into statistical learning ability as an individual difference (Hunt & Aslin, 2001; Kaufman et al., 2010; Schvaneveldt & Gómez, 1998), Kidd (2012) used the Serial Reaction Time (SRT) task, which is a common measure of procedural memory also employed in the current study. Nevertheless, questions remain to what extent there is an overlap between statistical learning, procedural memory, and indeed implicit learning constructs (see Kóbor et al., 2018; Monaghan et al., 2019).

3.2 The Current Study

The current study focuses on whether learners' ability to acquire an artificial language under cross-situational learning conditions is durable over time, and whether it is affected by individual differences in four memory systems that have been associated with language learning: phonological short-term memory, working memory capacity, declarative memory, and procedural memory. In doing so, it may shed light on the nature of the underlying mechanisms of cross-situational learning, and also add to the growing body of research that suggests it may be possible to explain language learning under immersion conditions in adults through a general-purpose, cross-situational statistical learning mechanism. In addition, investigating whether learning is durable is important from a methodological perspective, with possible implications for study design. Finally, the current study may also help refine theories of how memory

systems interact with language learning (e.g., Ullman, 2004, Ullman & Lovelett, 2016). Understanding what makes some adults better learners than others is essential to any theory of language learning.

We therefore report results of an experiment in which language learners were exposed to an artificial language under cross-situational learning conditions. We manipulated cross-situational scenes in order to test participants on their learning of nouns, verbs, adjectives, case markers, and word order, both during exposure and on a delayed test 24-hours later. Participants were also tested on five cognitive individual difference measures.

We predicted that, similar to Rebuschat et al. (2021), verbs and word order would be learned first, followed by nouns, adjectives, and case markers, although, given the short duration of the learning paradigm, it was possible that the latter may not be learned at all (e.g., DeKeyser, 2005). Based on the findings of studies investigating cross-situational learning (Vlach & Sandhofer, 2014) and statistical learning (Durrant et al., 2011; Gómez et al., 2006; Kim et al., 2009; and Vuong et al., 2016), we predicted that learning would be maintained after 24 hours in all aspects of the artificial language. Regarding the coherence or distinctiveness of vocabulary and grammar learning, we had two competing hypotheses. If there is a clear distinction between grammar and vocabulary, we expected learning of word order and case markers to be related on the one hand, and learning of nouns, verbs, and adjectives to be independently inter-related. This would mean that two underlying components may effectively describe learning of different aspects of the artificial language. Alternatively, if grammar and vocabulary share the same learning mechanisms, as is postulated in single-system models, we expected to see inter-relations between all aspects of the language. Finally, with regards to individual differences, while a majority of studies into working memory capacity and phonological short-term memory and

language learning suggest we could make firm predictions for their role in both vocabulary and grammar learning (see e.g., Baddeley et al., 1988; Mackey et al., 2002; Sagarra & Herschensohn, 2010), two factors may diminish their influence. Firstly, it remains a possibility that other cognitive measures (specifically procedural memory and declarative memory) may mediate the effects of phonological short-term memory and working memory capacity, thus better predicting grammar learning (Brooks & Kempe, 2013). If this were the case, we would not find such a strong link between the two working memory measures and vocabulary and grammar learning. Secondly, if working memory does not play such a strong role in incidental learning conditions as it does under explicit conditions, then that too may reduce the influence of phonological short-term memory and working memory capacity in this study. Based on Ullman's (2004) declarative/procedural model, we expected declarative memory to predict success during the early stages of acquisition. However, we left open the possibility that procedural memory might be a predictor for the more grammatical aspects, particularly under the more incidental learning conditions of cross-situational learning.

3.3 Method

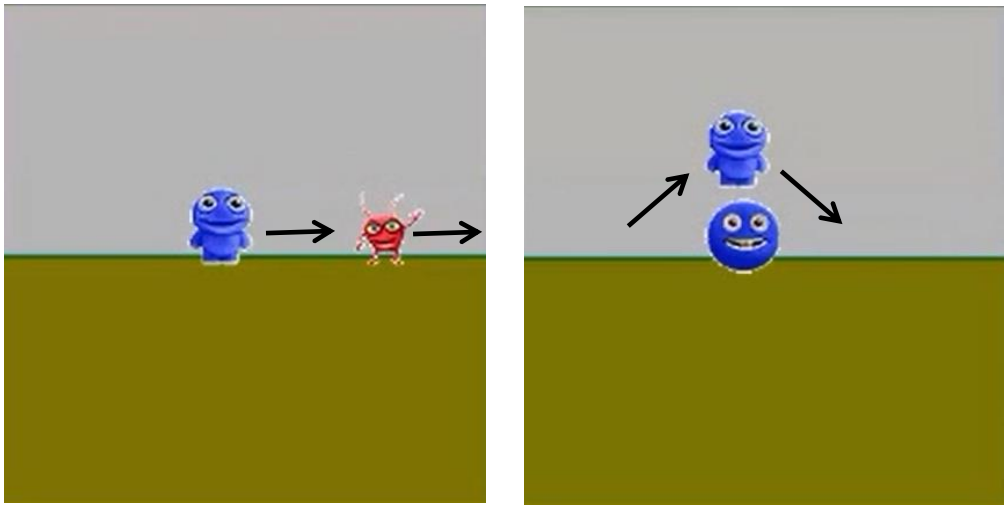
3.3.1 Participants

Sixty-four native speakers of English (47 women, 17 men) were randomly assigned to two exposure conditions (*massed* vs. *distributed*, each n = 32). These conditions varied in terms of whether there were three 20-minute pauses between blocks of exposure on an artificial language learning task.[1] Participants were students or graduates of the University of Central Lancashire or Lancaster University, both located in the North West of England. The mean age was 26.0 years (*SD* = 7.1). None of the participants had previously studied any verb-final

[1] These conditions did not exert any effect on performance and therefore will not be considered further.

languages. Four participants reported being bilingual in English and either Spanish, Punjabi, Thai, or Guajarati. Thirty-three participants reported knowing at least one other language to an intermediate-level of proficiency or above: Spanish (13), French (11), German (5), Polish (1), Italian (2), Portuguese (1), Chinese (2) and Malay (1). Twenty-two had previous experience of studying a case-marked language. Fifteen participants were studying or had studied in linguistics or for a language-related degree. Participants in the massed group received 20 GBP and participants in the distributed group received 28 GBP. The difference was due to the extra time involved in the distributed condition. Power analyses of the effect sizes found in Rebuschat et al. (2021) using G*Power demonstrated that 64 participants were sufficient to achieve .99 power for demonstrating learning of syntax, nouns, and verbs, .56 power for adjectives, and .05 power for case marker words (which are learned only under certain conditions). Despite the low power for case markers, we decided to include them in the current study for completeness. All recruitment was carried out in accordance with the ethical guidelines of Lancaster University and the University of Central Lancashire (see Appendix I).

Figure 4. Screenshot of the Cross-situational Learning Exposure Task



Note. Participants see two dynamic scenes and hear a sentence (e.g., "Hagal chilad tha garshal sumbad noo thislin"). Their task is to decide which scene the sentence refers to. The arrows indicate the direction of movement of the aliens and are for reference only, not appearing on the screen.

3.3.2 Materials

The artificial language developed by Rebuschat et al. (2021) was used in this experiment (see also Walker et al., 2017). The lexicon consisted of 16 pseudowords taken from Monaghan and Mattock (2012; see Appendix A). Fourteen bisyllabic pseudowords were content words: Eight nouns, four verbs, and two adjectives. Two monosyllabic pseudowords served as function words that reliably indicated whether the preceding noun referred to the subject or the object of the sentence. The words were recorded by a female native speaker of British English who was instructed to produce the words in a monotone voice.

In terms of syntax, the artificial language was based on Japanese. Sentences could either be SOV or OSV, i.e., verbs had to be placed in final position, but the order of subject and object

noun phrases (NPs) was free. NPs had to contain a noun as its head and a post-nominal case marker that indicated whether the preceding noun was the agent or the patient of the action. Adjectives were optional and only occurred in half the NPs. When adjectives were present, they occurred pre-nominally. The syntactic patterns used in the experiment can be found in Appendix B.

Eight alien cartoon characters served as referents for the language (Appendix C). The aliens could either appear in red or blue and were depicted performing one of four actions (hiding, jumping, lifting, pushing) in dynamic scenes generated by E-Prime (version 2.0). Figure 4 shows a sample screen shot, containing the target scene and a distractor scene. Each noun referred to an alien, adjectives referred to the colour of aliens, and verbs referred to their actions. Six different versions of word-referent mappings were randomly generated to control for preferences in associating certain sounds to objects, motions, or colours. Since adjectives were optional, sentences were between five and seven pseudowords in length.

3.3.2.1 Individual Difference Measures.

Phonological Short-Term Memory. Phonological short-term memory capacity was measured by means of Gathercole and Baddeley's (1996) nonword repetition test (NRT). This task required participants to listen to pseudowords of different lengths and to repeat each word exactly as they heard it. Answers were scored as described in Gathercole (1995), i.e., responses were either considered correct (1 point) or incorrect (0 points), depending on whether all phonemes had been reproduced correctly or not. Cronbach's alpha is .80.

Working Memory Capacity. The storage and processing function of working memory was measured by means of the Automated Operation Span (Aospan) Task by Unsworth et al.

(2005). During this computerised test, participants were presented with letters interspersed with maths problems, which they were required to solve while keeping the letters in memory at the same time. All letters were randomly selected for each subject from an array of twelve options (F, H, J, K, L, N, P, Q, R, S, T, and Y). The math problems always followed the pattern 'x multiplied/divided by y +/- z =n' (e.g., (16 / 2) + 3 = 11) and required the participant to choose whether the answer was true or false. The Aospan score was the sum of the number of letters that a participant was able to recall across blocks of increasing difficulty. Following Unsworth et al., an accuracy criterion of a minimum of 85% for the maths problems was set, which resulted in 10 participants being excluded. Unsworth et al. report a Cronbach's alpha of .78.

Declarative Memory. Visual declarative memory capacity was assessed by means of the Continuous Visual Memory Test (CVMT; Trahan & Larrabee, 1983). During the task, participants saw black-and-white drawings of complex figures, presented in succession, and were required to indicate if they had seen each picture before (i.e., if it was an "old" picture) or not (i.e., if it was "new"). D-prime scores were computed from the responses. The CVMT has a split-half reliability of .80. Verbal declarative memory capacity was assessed by means of the MLAT-V, a paired-associates test from the Modern Language Aptitude Test (Carroll & Sapon, 1959) that requires participants to rapidly learn 24 pseudo-Kurdish words and their English translations. Scores were calculated with one point for every correctly chosen item, for a maximum score of 24. Carroll and Sapon reported a split-half reliability of between .92 and .97.

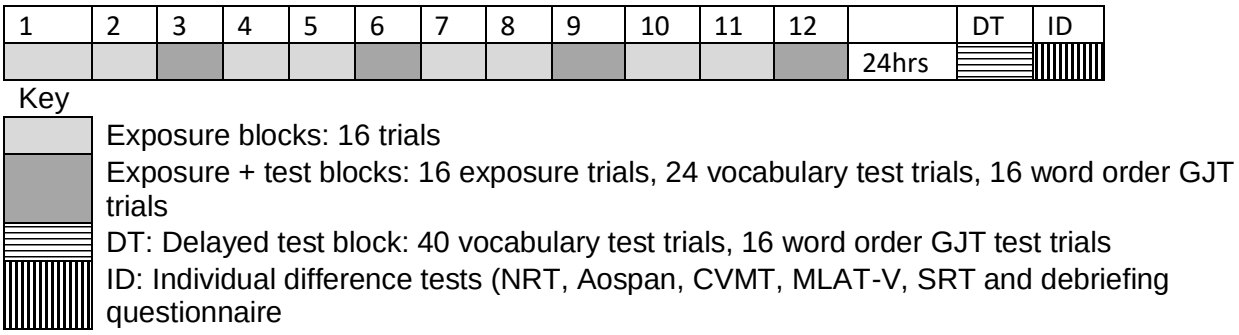
Procedural Memory. Procedural memory was measured through the Serial Reaction Time (SRT) task (Nissen & Bullemer, 1987, adapted by Lum et al., 2010). In this task, participants see a smiley face appear in one of four positions (top, right, bottom, left) on a computer screen. The task is to press the corresponding button (top, right, bottom or left) on a

game-pad controller as quickly and accurately as possible. Unbeknownst to participants, the location of the target is determined by a 10-item sequence (bottom, top, right, left, right, top, bottom, right, top, left). Participants repeat the same 10-item sequence over five blocks. On the sixth block, a pseudorandom sequence is introduced. In order to compute a procedural memory score, the median RT score of block five is subtracted from that of block six, with a higher, positive score indicating procedural learning. If participants implicitly start to learn the 10-item sequence, RTs are expected to slowly decrease from block 1 to 5 (structured sequences) and then increase in block 6 (random sequences). If, on the other hand, no procedural learning takes place, then median RTs for the sequenced block and the random block should be similar (Siegert et al., 2006).

3.3.3 Procedure

Participants were trained and tested on the artificial language over two days. On the day 1, participants completed twelve blocks of exposure to the cross-situational learning task, four blocks of which also tested vocabulary and word order. Blocks 1, 2, 4, 5, 7, 8, 10 and 11 were pure exposure blocks with 16 trials each. In blocks 3, 6, 9 and 12, intermingled with the 16 exposure sentences were 24 vocabulary test trials (eight each of noun and case marker test trials and four each of verb and adjective test trials). These four blocks then also included 16 trials of a grammaticality judgment task (GJT) to test participants' acquisition of the basic word order of the sentence (See Figure 5). No sentence-scene combination was repeated across the experiment.

Figure 5. Study 1 Research Design



Note. Participants were exposed to 192 sentences over 12 exposure blocks. Within every third exposure block were intermingled 24 vocabulary test trials and 16 word order GJT trials. The delayed test was administered after 24 hours, together with five individual difference tests and a debriefing questionnaire.

Twenty-four hours later, on day 2, participants returned to the lab to complete a fifth (delayed) test of vocabulary and word order. In this delayed test, there were 40 vocabulary test trials, consisting of eight each of noun, verb, and adjective test trials, 16 case marker test trials and 16 word order test trials. Presentation order of trials within each block was randomised but all participants completed blocks in the same sequence. Day 2 comprised the final block of vocabulary and word order test trials, five cognitive tests designed to measure individual differences in memory systems, and a debriefing questionnaire (see Appendix D). The five individual difference tests were administered in a randomly-assigned order. Exposure and testing on day 1 took between 90 and 120 minutes, and the delayed test, individual difference measures, and debriefing questionnaire on day 2 took around 60 minutes.

3.3.3.1 Exposure Trials.

For exposure, there were 12 cross-situational learning blocks with 16 exposure trials each. In each trial, participants were instructed to observe two dynamic scenes on the screen and listen to an artificial language sentence played over headphones. Their task was to decide, as quickly and accurately as possible, which scene the sentence referred to (see Figure 1). Participants received no feedback regarding accuracy. Within each block, each alien and action occurred an equal number of times; half the utterances in each block were SOV, the other half OSV. The locations of target and distractor scenes (left or right side of screen) were counterbalanced. In the distractor scene, no actions were the same as in the target scene, and the aliens and their colours were randomly selected.

3.3.3.2 Vocabulary Test Trials.

In order to make it less likely that participants would know they were being tested, test trials for each lexical category were intermingled in every third cross-situational learning block using the same cross-situational learning task as the exposure trials, but with one exception. The participant saw the two scenes, heard the sentence in the artificial ('alien') language, and was asked to select the scene to which the sentence referred. However, the target and distractor scenes were exactly the same apart from one piece of information: In noun testing trials, the target and distractor scenes were identical except for one of the aliens; in verb testing trials, only the scenes' actions differed; in adjective testing trials, one of the aliens' colour was changed; and in marker word testing trials, the two scenes depicted the same aliens performing the same actions but with agent-patient roles swapped. No feedback was provided on response accuracy.

3.3.3.4 Word Order Test Trials.

The acquisition of word order was tested by means of a grammaticality judgment task at the end of every third block of exposure. In this task, participants were told that they would see only one scene and hear a sentence spoken by another alien from a very different planet who was also learning the new language. Their task was to listen carefully and decide if the sentence sounded "good" or "bad" in relation to the alien language. The pseudowords always matched the scene but only half the sentences were grammatical, following the licensed SOV or OSV word order. The other half of the sentences contained syntactic violations (*VSO, *VOS, *OVS, *SVO). That is, the task tested sensitivity to correct sequencing of phrases (noun phrase, noun phrase, verb) rather than sequences within phrases (e.g., adjective, noun, marker word, within the noun phase). None of the grammaticality judgment task sentences occurred during exposure trials. Again, no feedback was provided.

3.4 Results

3.4.1 Performance on Exposure Trials

In order to ascertain when learning had taken place during the exposure blocks, a one-sample t-test was conducted to compare the mean scores for each block to a chance score of .5. Table 2 gives a summary of the findings for the one-sample t-test and the descriptive statistics. Participants performed significantly above chance from block two (M = .57, SD = .18) onwards, 95% CI [.53 to .62], $t(63) = 3.27$, $p = .002$. Thus, 32 trials of exposure (without feedback) were enough to lead to above-chance performance in the cross-situational learning task.

Table 2. Summary of Descriptive Statistics and One-Sample T-Tests on Mean Scores for Each Block of Exposure

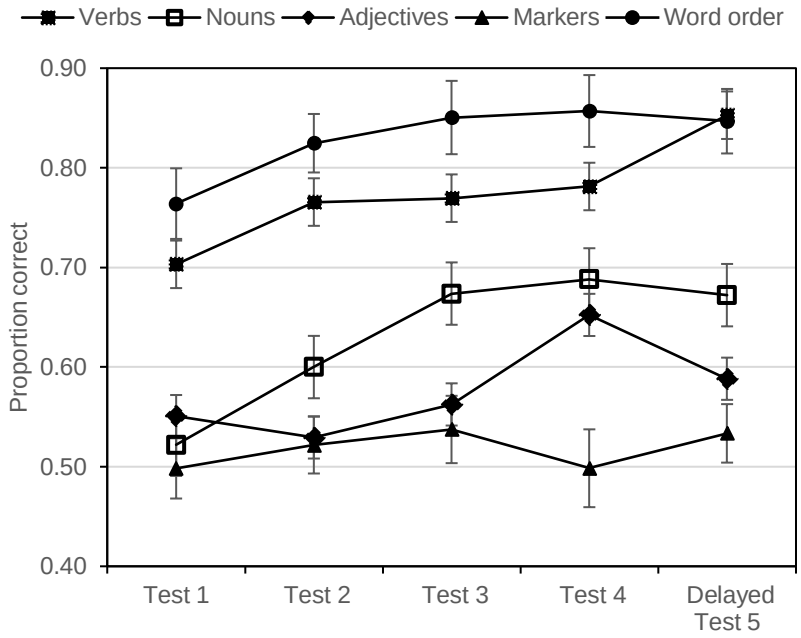
Exposure block			95% CI	95% CI			*Sig.. (2-*
	M	*SD*	*Lower*	*Upper*	*T*	*df*	*tailed)*
Exposure block 1	.53	.15	.49	.57	1.498	63	.14
Exposure block 2	.57	.18	.53	.62	3.265	63	.002
Exposure block 3	.70	.18	.66	.74	8.883	63	<.001
Exposure block 4	.76	.19	.71	.80	10.812	63	<.001
Exposure block 5	.75	.21	.70	.81	9.697	63	<.001
Exposure block 6	.79	.20	.74	.84	11.900	63	<.001
Exposure block 7	.79	.22	.74	.85	10.622	63	<.001
Exposure block 8	.82	.22	.76	.88	11.528	63	<.001
Exposure block 9	.84	.19	.80	.89	14.170	63	<.001
Exposure block 10	.83	.23	.78	.89	11.803	63	<.001
Exposure block 11	.85	.21	.80	.90	13.556	63	<.001
Exposure block 12	.85	.20	.80	.90	14.078	63	<.001

3.4.2 Performance on Test Trials

As the grammaticality judgment test was a two-way forced choice between grammatical and ungrammatical sentences, we first conducted a signal detection analysis to check that accuracy scores were an acceptable reflection of discrimination, rather than response bias. In order to do this, d' (prime) and C (bias) scores were calculated for each of the word order tests 1-4. They were then entered into repeated measures ANOVAs. Greenhouse-Geisser corrected results, showed that d' was discriminative, $F(2.270, 143) = 14.585$, $p < 0.001$, but not C, $F(2.168, 136.5) = 1.856$, $p = 0.157$. Accuracy scores were thus taken as a true reflection of discrimination and were used for the further analyses. Employing accuracy rather than d' ensures that grammaticality judgement and vocabulary tests are more comparable.

We next looked at the effects of learning syntactic word order, nouns, verbs, adjectives and case markers over the two days by conducting repeated measures ANOVAs with reverse Helmert contrasts on accuracy for learning each aspect of the language across tests 1 to 5. Reverse Helmert contrasts indicate the point at which a significant step-change in learning occurs from previous to subsequent blocks. For a summary of the reverse Helmert contrasts see Appendix E. Greenhouse-Geisser correction was applied when the assumption of Sphericity was not met. The results are shown in Figure 6 and Table 3. Learning of word order, nouns, and verbs all improved significantly across the five tests.

Figure 6. Proportion of Correct Trials across the Five Tests



Note. Tests 1 to 4 were completed on day 1, with test 5 administered after a 24-hr delay.

The error bars represent standard errors of the mean.

Table 3. Summary of Repeated Measures ANOVA over Tests 1 to 5 Showing Effect for Test Block

Effect	*Df*	*F*	*P*	η_p^2
Word order	2.812, 177.170	7.20	<.001	.10
Noun	4, 252	12.4	<.001	.17
Verb	4, 252	5.27	<.001	.077
Adjective	4, 252	2.30	.059	.035
Case marker	4, 252	.86	.59	.013

For word order, participants' test scores improved significantly from each test to the next, including from day 1 to day 2, $F(1, 63) = 4.136$, $p = .046$, $\eta_p^2 = .062$. A one-sample t-

test shows that word-order scores were significantly above chance from block 1 onwards (See Appendix F for one-sample t-tests; M = .76, SD = .19, 95% CI [.72 to .81], $t(63) = 10.9$, $p <$ .001).

Accuracy for nouns also improved significantly between each of the first four tests on day 1 despite a significant drop in scores from day 1 to day 2, F (1, 63) = 5.650, p = .021, η_p^2 = .082. One-sample t-tests show that results were significantly above chance from noun test 2 onwards, M = .60, SD = .19, 95% CI [.55 to .65], $t(63) = 4.16$, $p < .001$, including for noun test 5, M = .68, SD = .25, 95% CI [.47 to .83], $t(63) = 5.69$, $p <$ .001.

Accuracy for verbs significantly improved across the first four tests, although a one-sample t-test shows that this is because participants learned verbs early and were already significantly above chance by verb test 1, M = .70, SD = .25, 95% CI [.63 to .76], $t(63) = 6.16$, $p < .001$. Interestingly, there was a significant improvement in verb scores from day 1 to day 2, F (1, 63) = 19.028, p < .001, η_p^2 = .23.

Adjectives showed no significant improvement over the five tests, although scores for adjective test 4 were significantly higher than mean scores of the previous three adjective tests, F (1, 63) = 8.401, p = .005, η_p^2 = .118. No significant change in accuracy was found for adjectives from day 1 to day 2, with one-sample t-tests confirming that both adjective test 4, M = .64, SD = .27), 95% CI [.58 to .71], $t(63) = 4.22$, $p < .001$, and test 5, M = .55, SD = .22), 95% CI [.53 to .65], $t(63) = 2.92$, $p = .005$, were significantly above chance.

For case markers, no significant step change in improvement over time occurred. Test 5, was, however, just significantly above chance, M = .54, SD = .16), 95% CI [.50 to .58], $t(63) = 2.06$, $p = .043$. In the ANOVA, there were no significant main effects or interactions with massed/distributed exposure condition, with all $F < 2.3$ and all $p > .13$ in all cases.

There were no significant differences in test results for linguistics or language students, those who had at least one other language to intermediate level, or for knowledge of case marked languages. For a summary of the results by language profile see Appendix G.

The debriefing questionnaire offered insights into the extent to which learning was verbalizable, as well as increased detail regarding what participants learned from the study. For word-order, 47 participants (73%) could identify the basic noun phrase–noun phrase–verb syntax, but only 17 participants (26.6%) correctly identified the exact word order of the noun phrase (adjective–noun–marker). Regarding case markers, only 14 participants (21.9%) could correctly identify that they referred to the agent and patient of the sentence. In terms of the lexical item that participants noticed first, 38 (59.4%) reported that they first noticed verbs, and only 15 (23.4%) reported noticing nouns first.

3.4.3 Determining Relations Between Learning Different Information Types

In order to determine which underlying component(s) drove performance in the task, that is, whether learning was independent or interdependent for different types of information, we conducted an exploratory principal components analysis on test performance for word order, nouns, verbs, adjectives, and case markers in tests 1 to 4 combined and for the final, delayed test. For both tests 1 to 4 combined and test 5, there were two components with eigenvalues greater than 1, and the loadings of the individual tests on these components, with varimax rotation, showed a simple solution (i.e., each test loaded > 0.4 on only one component). Due to this simple structure, we did not separately relate the learning of types of information to the individual difference measures. The components and their loadings are shown in Table 4 and Table 5.

Table 4. Loadings of the Five Tests on the Two Principal Components for Tests 1-4 Combined

Tests 1-4	First component	Second component
Noun	**.626**	.145
Adjective	**.763**	.212
Case marker	**.692**	-.210
Verb	-.034	**.848**
Word order	.140	**.690**

Table 5. Loadings of the Five Delayed Tests on the Two Principal Components for Test 5

Test 5	First component	Second component
Noun	**.778**	.104
Adjective	**.769**	.034
Case marker	**.604**	.081
Verb	.322	**.718**
Word order	-.090	**.873**

For tests 1 to 4 combined, the first component related to learning nouns, case markers, and adjectives, and the second component related to learning word order and verbs. This indicated that performance across the five information types was effectively explained by two aspects of the data: The first relates to learning the vocabulary items of nouns and adjectives and how the marker words affected the role of the adjective-noun phrases. The second indicated a close relation between learning the identities of verbs and learning that the word order of sentences was verb-final. We return to this point in the Discussion. For the delayed

test 5, a similar picture emerged: The first component related to learning nouns, adjectives, and marker words, and the second component related to learning word order and verbs.

3.4.4 Individual Difference Measures

Scores for measures of the five individual cognitive measures were converted into z-scores (see Table 6).

Table 6. Descriptive Statistics for Individual Difference Measures

ID measure	*M*	*SD*	*SE*	Range
NRT	21.4	2.68	0.34	12
Aospan	38.0	19.4	2.44	75
CVMT	1.80	0.55	0.07	2.94
MLAT-V	16.8	4.89	0.61	19
SRT	54.2	35.6	4.45	185

We conducted a series of stepwise linear regressions to determine the relations between the individual difference measures and the two components of learning the language derived from the principal component analysis. All five individual difference measures were included in the same level. Again, we distinguished performance on day 1 (i.e., tests 1-4 combined) and performance on day 2 (test 5) as the dependent variables in separate regressions, and the five individual difference measures (NRT, Aospan, CVMT, MLAT-V and SRT) as the independent variables. No predicting variable was found for test 5: word order, verbs, and case markers. Table 7 shows the predicting variables for each of the principal components for tests 1-4 and test 5.

Table 7. Summary of Step-Wise Linear Regression with Principal Component Analysis Variable Scores for Lexical and Word Order Tests 1-4 Combined and Test 5 as the Dependent Variables and the Five ID Measures (NRT, Aospan, CVMT, MLAT-V and SRT) as the Independent Variables

Component		*b*	*SE b*	*B*
Test 1-4: component 2 (word order and verbs)	Constant	.01	.12	
	MLAT-V;	.25	.16	.25**
Test 1-4: component 1 (nouns, adjectives and markers)	Constant	-.011	.12	
	SRT	.35	0.12	0.35***
Test 5: component 1 (nouns, adjectives and markers)	Constant	.009	.12	
	MLAT-V	.36	.12	.36**

3.5 Discussion

In this study, we replicated and extended Rebuschat et al. (2021) by investigating whether, and, if so, in which order, adult learners could acquire the syntax and vocabulary of a novel language via cross-situational learning, without feedback, and without any explicit instruction about the structure of the language or its vocabulary. We then further explored this paradigm further by implementing a 24-hours delayed posttest to determine whether any acquired knowledge had been maintained. Furthermore, we investigated how learning of syntax and vocabulary cohered, and which cognitive individual differences (particularly working memory capacity, phonological short-term memory, declarative memory, and procedural memory) affected learning of different aspects of the artificial language.

3.5.1 Learning Under Cross-Situational Conditions

Our results indicated that adult learners can rapidly acquire both vocabulary and certain aspects of syntax of the language simultaneously, in line with Rebuschat et al. (2021).

While this study did not investigate the mechanisms underlying the learning, our findings are consistent with the use of a combination of statistical learning mechanisms (Yu & Smith, 2007), syntactic and semantic bootstrapping (Abend et al., 2017; Gleitman, 1990), and a propose-but-verify procedure (Trueswell et al., 2013), with initial learning occurring implicitly (consistent with associative learning mechanisms) and then top-down knowledge interacting more explicitly and working in unison with unconscious learning (consistent with strategic approaches to learning). Replicating the findings of Rebuschat et al. (2021), verbs and basic word order were learned to a level above chance first, followed by nouns[2], then adjectives, and finally case markers (while case markers did reach a significant level above chance for test 5, the lack of main effects on the repeated measures ANOVA suggests we ought to conclude that learning did not occur in their case). The fact that verbs were learned to a level above chance before nouns may relate to the increased salience of the sentence-final verbs (Fernald et al., 1992; Shoemaker & Rast, 2013), consistent with recency effects in sequence processing that Freudenthal et al. (2007) demonstrated can explain a variety of morphological and grammatical effects in language learning, and Jones and Rowland (2017) suggested can explain phonotactic effects on word learning. However, it is possible that the verb-final position and its immovability provided an extra recency advantage that is not generally available in language learning and L1 acquisition in particular, by increasing the likelihood of verbs and their position in the sentence reaching the level of awareness, and thus providing opportunities for more metacognitive strategy use. Another possible contributor to the order of acquisition effects observed in this study is that the artificial language included eight nouns but only four verbs, potentially making the latter easier to learn. Yet, despite these caveats, the similarity to the order of acquisition in verb-dominant languages (Choi & Gopnik, 1995; Tardif, 1996) raises the question of whether and to what

[2] We do not know if participants learned proper or common nouns. However, we know that they learned the labels for the aliens and that these are nouns.

extent the findings of this particular study can be generalised to first language acquisition. While all our participants were adults who already possessed syntax, vocabulary, and often metalinguistic knowledge from their first languages, we contend that the underlying mechanisms of cross-situational learning are similar for both first and subsequent languages under immersion conditions, albeit with differing levels of pre-existing knowledge and cognitive development affecting the relative roles of explicit and implicit learning.

3.5.2 The Durability of Cross-Situational Learning

We built upon the previous study to investigate whether learning under cross-situational learning conditions is durable over time. Importantly, the results showed that the learning effects can be retained overnight, and performance actually improved with tests for verbs, word order, and case markers, albeit for the latter non-significantly. This is an important methodological observation as the majority of studies in cross-situational learning do not have a delayed posttest, which means that it is unclear whether the learning is robust. By including a 24-hour delayed posttest, we show that learning is indeed robust and that this applies to words and syntax. It is recommended, therefore, that future studies into cross-situational learning include delayed posttests to show that learning is robust, and to catch any learning effects brought on through consolidation.

Coherence of Syntax and Vocabulary

Regarding the coherence of learning of syntax and vocabulary, we found that acquisition of word order and verb learning were interdependent. As to why word order and verbs cohered, this could at least be partially explained by the nature and simplicity of the word order test. The grammaticality judgment task could be completed successfully by noticing that in the ungrammatical sentences (*OVS, *VSO, *SVO, *VOS) the last word of the sentence was mono-syllabic (the post-nominal case marker) rather than bi-syllabic (all the

content words but particularly verbs). In the debriefing questionnaire, a large majority of participants reported that the first thing they explicitly noticed was that the final word was an action. Questionnaires are only able to capture verbalizable, explicit knowledge, so it is impossible to tell from this instrument what was happening at a subconscious, implicit level. However, taken together with the results showing a coherence of learning of word order and verbs, and consistent with the findings of Khoe et al. (2019) that ambiguous learning situations rely on associative learning rather than hypothesis-testing strategies, this tentatively suggests that participants first learned one referent for the final position, and therefore its word category (captured in the grammaticality judgment results), before going on to learn the referents for the other actions. In addition, we found that nouns, adjectives, and case markers were also interdependent but were acquired somewhat independently of verbs and word order. These lexical categories comprise the noun phrase, and it is therefore perhaps unsurprising that success with adjectives and case markers corresponds to success with nouns, on which their meanings and syntactic roles depend. In sum, these results support a view of language in which the syntactic knowledge associated with case markers begins to develop only after the syntactic roles and semantic meanings of a core vocabulary of content words has been learned (Bannard et al., 2009).

3.5.3 Individual Differences in Cross-Situational Learning

The final objective of this study was to investigate the role of short and long-term memory systems (phonological short-term memory, working memory capacity, declarative memory, and procedural memory) in the acquisition of this artificial language under cross-situational learning conditions. Neither working memory capacity, as measured by the Aospan, nor phonological short-term memory, as measured by NRT, predicted success on the lexical test scores. This mirrors recent findings into incidental learning conditions (Hamrick,

2015; Tagarelli et al., 2015). It is possible that working memory capacity is utilised mainly when language is learned under explicit conditions.

For tests 1 to 4 combined, regressions of the principal component factor scores revealed that component 1, which included nouns, adjectives, and case markers, was predicted by SRT, a measure used to assess procedural memory, while component 2, which included verbs and word order, was predicted by MLAT-V, a measure used to assess declarative memory. On the surface these results do not tally with proposals that associate vocabulary learning primarily with declarative memory, and grammar learning with procedural memory (Ullman, 2004), nor with declarative memory being responsible for all the early stages of learning including grammar (Hamrick, 2015; Morgan-Short et al., 2014; Ullman, 2016). However, one possible interpretation of these results involves the amount of attentional resources dedicated to each lexical category at different times during the study. As mentioned earlier, Ullman and Lovelett (2016) state that altering the learning conditions may force more of a reliance on either declarative or procedural memory. While no deliberate manipulation occurred in this study, it is possible that this was a by-product of the exposure task and lexical tests used. In order to complete the exposure blocks on the first day, participants may have explicitly and strategically focused on verbs (and associated word order), and so success with these would be predicted by declarative memory. Nouns, adjectives, and case markers, on the other hand, would have received less explicit attention and thus learned more incidentally. This might explain why success in these lexical categories, therefore, was predicted by procedural memory. On the second day, with verbs and word order reaching ceiling effects, participants could shift their attention to nouns, adjectives, and case markers. As a result, success with these categories was now predicted by declarative memory. While we acknowledge that there is a possibility that this is a consequence of the construction of this particular artificial language, we believe that it is

more analogous to the shifting of attention at different stages of real language learning. A further study in which attention is directed towards certain lexical categories might shed light on this aspect of the declarative/procedural model.

An alternative explanation for these results is that acquiring the lexical categories which comprised the noun phrase required more pattern learning than for verbs and for completion of the grammaticality judgment task, which tested knowledge of basic word order. Adjectives were optional in our artificial language; nouns could therefore adopt a number of different positions in the sentence, and case markers further influenced the position of nouns and adjectives. Thus, in this interpretation, success with the components of the noun phrase was predicted by procedural memory. Performance on verb tests and the grammaticality judgment task, on the other hand, only required attention to the last word in the sentence, and so their syntactic roles in the sentence were perhaps less important than their semantic meanings. Hence, success with this component was predicted by declarative memory. Ullman and Lovelett (2016) do indeed assert that procedural memory is responsible for the more rule-governed aspects of vocabulary acquisition. Therefore, it is possible that in the early stages of the acquisition process under cross-situational learning conditions, procedural memory plays an important part in determining the syntactic roles of vocabulary (Evans et al., 2009). Turning to the 24-hour delayed test 5, component 1, which contained adjectives, nouns, and markers, was predicted by MLAT-V, the measure of declarative memory. Follow-up analyses from debriefing questionnaires on the awareness of rule knowledge of word order and the function of case markers suggest that declarative memory may play a role on delayed tests when learners become aware of the rules, that is, when rules become explicit and verbalizable. Component 2, which contained verbs and word order, did not have a predictor. This may have been affected by a ceiling effect on scores for verbs and the grammaticality judgment task. While both these explanations are plausible, neither is

fully satisfactory, and nor do either of them account for the early-stage dominance of declarative memory (Ullman, 2016). Further research is needed before the relationships between declarative memory, procedural memory, early vs. late stages, attention vs. less attention, and arbitrary vs. rule-governed items are fully understood.

One issue with the research tools used in the current study is whether the SRT task tested statistical learning or procedural memory. In order to best tap into procedural memory, as a long-term memory system, one solution would be to include a delayed posttest on the SRT. This has been done in several recent studies (Desmottes et al., 2016; Desmottes et al., 2017; Hedenius et al., 2011) into specific language impairment (SLI), which has been hypothetically linked to a deficit in procedural memory (Ullman & Pierpont, 2005). It is possible, therefore, that individual differences for a delayed posttest for SRT would be better predictors for performance on delayed tests for nouns, adjectives and case markers.

3.6 Conclusion

Our study confirms that it is possible for adults to learn syntax and vocabulary simultaneously under cross-situational learning conditions, and that the order of acquisition follows verb-dominant language acquisition, but also that the learning effect persists after 24 hours. The durability aspect is important from a methodological perspective, as not only does this show that knowledge is robust but also that, without a delayed test, learning can be in some cases underreported. Moreover, the patterns of results we found for this verb-final language in our experimental paradigm did not neatly correspond with a distinction between grammar and vocabulary learning (e.g., Ullman, 2004), with word order being related to verb acquisition, and case marking being related to noun and adjective learning. Complex interactions between lexical categories and grammar do not appear to lend themselves to a clear distinction in acquisition of these sources of linguistic knowledge, nor in the individual differences that predicted them. Future studies into language learning in adults under cross-

situational learning conditions should continue to investigate individual differences and how they cohere, as it is a central question to both language learning theory and pedagogical applications such as computer assisted language learning design (Meurers et al., 2019).

A limitation to this study is the number of participants (n=64) compared to the number of analyses that were carried out. Caution should therefore be exercised before generalizing the findings. The extent to which the effects we observe here are due to participants' first language exposure or are generic for language learning, remains a matter of debate. In any case, this paradigm offers new opportunities to investigate cross-linguistic differences in early-stage acquisition of vocabulary and grammar (see Fedzechkina et al., 2016 for an overview), and the potential support and interference of different aspects of learning across related and unrelated languages.

In the following chapter 4, the data from study 1 is analysed with regard to the effect that massed and distributed practice schedules affecting the order of acquisition, the strength of learning, and retention and consolidation after 24 hours. Interactions between massed and distributed schedules and the five individual memory differences are also investigated.

Chapter 4: Cross-situational Learning under Massed and Distributed Schedules: Additional analysis of Study 1

4,1 Introduction

Chapter 3 included an article that was published in Language Learning (Walker et al., 2020). In that paper, I described study 1 and then analysed it with regard to whether a cross-situational learning paradigm could be used to learn an artificial language with nouns, verbs, adjectives and case markers; what the order of acquisition was; whether that learning was durable after 24 hours; and whether different aspects of the learning were interdependent and were predicted by individual memory difference measures. Results showed that not only was it possible to learn this artificial language through cross-situational language learning, an incidental learning process that provides no feedback to the learner, but that this learning was durable after 24 hours. Verbs and a basic word order were learned first, followed by nouns and adjectives, while case markers were not reliably learned. Verbs and basic word order were interdependent and predicted by MLAT-V, a measure of declarative memory. The parts of speech constituting the noun phrase were also interdependent and were predicted by SRT, a measure of procedural memory. However, the article that constituted chapter 3 did not include further analysis of the differences between massed and distributed groups' learning schedules. Therefore, in chapter 4, I will report an additional analysis of the data to investigate whether distributing the presentation and practice of the cross-situational learning task improved learning compared to massing it. I also reanalysed the data from the five individual memory difference measures (visual and verbal declarative memory, procedural memory, working memory capacity and phonological short-term memory) to determine whether they differentially predicted success with massed and distributed groups.

4.2 Review of Relevant Literature

A growing area of research interest in terms of L2 language learning under more naturalistic, incidental conditions is cross-situational language learning, which has been

hypothesised to account for early-stage learning under naturalistic, immersion conditions (Monaghan et al.; 2019; Monaghan et al., 2015; Rebuschat et al. 2021; Scott & Fisher, 2012; Smith & Yu, 2008; Yu & Smith, 2007). Therefore, experimental paradigms such as that used by Rebuschat et al. (2021) and study 1 in this thesis may be ideally placed for use in lab-based investigations into learning under immersion settings and into low-cost pedagogical interventions that might enhance learning and retention under these conditions. One such intervention, and the focus of the current study, is the temporal distribution of study.

The distributed practice effect, a benefit for spaced rather than massed practice, is one of the most robust findings in cognitive psychology (for reviews, see Cepeda et al., 2006; Delaney et al., 2010), found in many domains of learning (see Küpper-Tetzel, 2014), ages (e.g.,Kornell et al., 2010), and even species (e.g.,Menzel et al., 2001). One finding, at least with regards to paired-associate learning tasks, is that incidental learning conditions also demonstrate a distributed practice effect, albeit one that is weaker and requires a narrower optimal lag than intentional learning conditions (Janiszewski et al., 2003). However, in L2 language acquisition, while strong distributed practice effects have been found for vocabulary learned under intentional, rote-learning conditions (Bahrick et al., 1993; Bloom & Schuell, 1981; Küpper-Tetzel et al., 2014; Gerbier et al., 2015; Kornmeier et al., 2014), results have been more varied in in terms of distributed practice studies in both L2 contextual vocabulary learning, in which vocabulary is embedded within graded readers to be incidentally acquired (Elgort & Warren, 2014; Macis et al., 2021; Serrano & Huang, 2018; Webb & Chang, 2015) and L2 grammar (Bird, 2010; Miles, 2014; Rogers, 2015; Suzuki & DeKeyser, 2017a; Suzuki, 2017; Kasprowicz et al., 2019). To the best of my knowledge, only one study has so far investigated distributed practice under cross-situational learning conditions. Vlach et al. (2012) found that two-year olds were better able to abstract and generalise nonce nouns under cross-situational learning conditions when schedules were distributed compared to massed.

However, ISIs were in seconds and the RI was at 15mins, and only one noun was learned at a time. It has yet to be determined if distributing practice of cross-situational learning over longer periods and with multiple lexical categories confers a similar benefit and whether this translates to adults exposed to an L2.

An under-investigated area of distributed practice research is how individual differences in memory interact with massed and distributed schedules. While there have been studies investigating certain other individual differences, including age (Toppino & DiGeorge, 1984), language proficiency (Elgort & Warren, 2014)), IQ (Madsen, 1963), visual-spatial ability (Mumford et al., 1994), to the best of my knowledge only one study has looked at individual memory differences (although see Camp et al.,1996, for a study into distributed practice for patients with dementia). Suzuki and DeKeyser (2017b) investigated the learning of Japanese morphosyntax under intentional learning conditions with different lag schedules. They found that working memory capacity correlated with a shorter 1-day lags, while language analytic ability correlated with a longer 7-day lag. There is a surprising lack of research into other memory constructs (declarative memory, procedural memory, phonological short-term memory) and their potential interaction with distributed and massed schedules, particularly for L2 language learning under incidental learning conditions. Investigating the role that these memory constructs play in the distribution of cross-situational learning schedules may help our understanding of the underlying mechanisms of the distributed practice effect, cross-situational learning, and may help pave the way for a more individual difference-based learning schedules.

4.3 Research Questions

1) Does distributed practice (20-minute ISI) of an artificial language presented under cross-situational learning conditions result in better results on a delayed posttest (1-day RI) than massing presentations?

2) Do individual difference measures (visual and verbal declarative memory, procedural memory, phonological short-term memory and working memory capacity) predict success with the cross-situational learning task under massed or distributed learning conditions? Is there an interaction between individual difference measures and massed or distributed conditions?

4.4 Hypotheses

I predicted that there would be a distributed practice effect on the delayed 24-hour posttest for all lexical categories. Due to the wealth of research into distributed practice and rote-learning of vocabulary (e.g., Bahrick et al., 1993), together with the one study into vocabulary learned under cross-situational learning conditions (Vlach et al., 2012), I predicted that the "vocabulary" items (verbs, nouns and adjectives) would have a larger effect than the "grammatical" items (word order and case markers).

Following Suzuki and DeKeyser (2017), I predicted that working memory capacity as assessed by the Aospan would predict success on lexical items under massed learning conditions. I also hypothesised that procedural memory, as measured by the Serial Reaction Task, would predict success under distributed conditions during training due to the differences in the amount of offline consolidation occurring during the lag. Drawing on theories of desirable difficulty (Benjamin & Tullis, 2010), I predicted that declarative memory (verbal and visual), as measured by the MLAT-V and the CVMT, would predict success on the delayed 1-day RI posttest for the distributed schedule. The reason for this hypothesis is that the increased lag would raise the level of difficulty for the participants, creating a desirable or optimal level of difficulty for those with stronger declarative memories and an undesirable, sub-optimal level of difficulty for those with weaker declarative memories.

I did not have any firm hypotheses for how the other individual difference measures would interact with the massed and distributed learning conditions. These individual difference measures were thus included for exploratory purposes.

4.5 Method

As this chapter is a reanalysis of the data from study 1, a full description of the participants, the materials and the procedure can be found in chapter 3. However, to aid the reader, here I will provide a brief recap.

4.5.1 Participants

Sixty-four undergraduate and graduate participants (X female) were recruited from two universities in the northwest of England. None of the participants had studied a verb-final language such as Korean or Japanese.

4.5.2 Materials

The artificial language used in this experiment was taken from Rebuschat et al. (2021) and comprised of 16 pseudowords taken from Monaghan and Mattock (2012). Fourteen bisyllabic content words: eight nouns represented eight different alien cartoon characters, four verbs representing actions (hiding, jumping, lifting, pushing), and two adjectives depicting the colour of the aliens (red or blue); two monosyllabic pseudowords served as function words that indicated whether the preceding noun referred to the subject or the object of the sentence. Sentences could either be subject-object-verb (SOV) or object-subject-verb (OSV). NPs had to contain a noun as its head and a postnominal case marker that indicated whether the preceding noun was the agent or the patient of the action. Adjectives were optional and only occurred in half the NPs. When adjectives were present, they occurred pronominally.

Eight alien cartoon characters served as referents for the language. The aliens could either appear in red or blue and were depicted performing one of four actions (hiding, jumping, lifting, pushing) in dynamic scenes generated by E-Prime (version 2.0). Figure 1 shows a sample screen shot, containing the target scene and a distractor scene. Each noun referred to an alien, adjectives referred to the colour of aliens, and verbs referred to their actions. Six different versions of word-referent mappings were randomly generated to control for preferences in associating certain sounds to objects. For example;

	haagle	chelad	tha	goorshell	sumbark	noo	fisslin
gloss:	blue	Alien7	OBJECT	red	Alien5	SUBJECT	jumps

Red alien5 jumps over blue alien7

4.5.2.1 Individual Difference Measures. Five measures of different aspects of memory were taken. Verbal declarative memory was measured via the paired-associates test (part V) of the Modern Language Aptitude Test (Carroll & Sapon, 1959), in which participants are required to learn 24 pseudo-Kurdish words and their English translations. Visual declarative memory was assessed via the Continuous Visual Memory Test (CVMT; Trahan & Larrabee, 1983), in which participants were presented with a succession of complex figures and were required to choose whether the shape had been presented previously. Procedural memory was assessed by means of the Serial Reaction Time task (Nissen & Bullemer, 1987, adapted by Lum, Gelgic, & Conti-Ramsden, 2010), in which participants responded to a ten-item sequence of stimuli, with learning measured via reaction time drop offs after a pseudo-random sequence was introduced. Phonological short-term memory capacity was measured by means of Gathercole and Baddeley's (1996) Nonword Repetition Test (NRT), in which participants' ability to repeat and pronounce unfamiliar phonological sequences was assessed. Working memory capacity was measured via the Automated Operation Span (Aospan) task (Unsworth et al., 2005), in which participants

solved math problems while remembering a series of unrelated items. For a fuller description of the individual difference measures, see chapter 3.

4.5.3 Procedure

As described in chapter 3 (see section 3.3.3), participants completed the experiment over two days, with twelve blocks of exposure and testing on day one, and the delayed test and individual difference measures on day 2, 24 hours after the exposure blocks. Each exposure block consisted of 16 exposure trials. In every third block, in addition to the exposure trials, 24 vocabulary test trials and 16 word order grammaticality judgement test trials were administered. The massed condition group conducted all 12 blocks back-to-back while the distributed condition block had three 20-minute gaps after blocks 3, 6 and 9 in which they watched nature documentaries on mute. The ISI was selected based on the findings from Kornmeier et al. (2014) that 20 minutes was an optimal spacing for a 24-hour RI. The delayed posttest was administered after 24 hours and included 24 vocabulary test trials and 16 word order grammaticality judgement test trials. See Figure 7.

Figure 7. Study 1 Distributed Practice Research Design

Massed Group																	
			1	2	3	4	5	6	7	8	9	10	11	12		DT	ID
															24hr		
Distributed Group																	
1	2	3		4	5	6		7	8	9		10	11	12		DT	ID
			20min				20min				20min				24hr		

	Exposure blocks: 16 trials
	Exposure + test blocks: 16 exposure trials, 20 vocabulary test trials, 16 word order GJT trials
	DT: Delayed test block: 24 vocabulary test trials, 16 word order GJT test trials
	ID: Individual difference tests (NRT, Aospan, CVMT, MLAT-V, SRT

Note. Participants were split into two groups (massed and distributed). In the massed group, there were no gaps between the exposure blocks. In the distributed group, the 12 exposure blocks were separated by three 20-min gaps (ISI) after blocks 3, 6 and 9. The delayed test was administered after 24 hours.

4.5.3.1 Exposure Blocks. Each exposure block consisted of 16 exposure trials. In each trial, participants were instructed to observe two dynamic scenes playing simultaneously on a screen (E-Prime, version 2.0). In each scene the participant saw either a red or a blue alien performing an action on another red or blue alien. They also heard a sentence in the alien language. Participants were instructed to choose, as quickly and accurately as they could, the scene which corresponded to the sentence. No feedback was given.

4.5.3.2 Vocabulary Test Trials. Vocabulary test trials were identical to exposure trials with one exception. The target and distractor scenes were exactly the same apart from the piece of information being tested. In noun testing trials, the target and distractor scenes were identical except for one of the aliens; in verb testing trials, only the scenes' actions differed; in adjective testing trials, one of the aliens' colour was changed; and in marker word testing trials, the two scenes depicted the same aliens performing the same actions but with agent-patient roles swapped. Again, no feedback was provided on response accuracy.

4.5.3.3 Word Order Test Trials. In every third exposure block, after the vocabulary test trials, a grammaticality judgement test was administered to test the acquisition of word order. In this task, participants saw one scene and heard one sentence. Participants were instructed to listen carefully and decide if the sentence sounded "good" or "funny" in relation to the alien language. The pseudowords always matched the scene but only half the sentences were grammatical, following the licensed SOV or OSV word order.

4.6 Results

4.6.1 Training Blocks and Tests 1-4

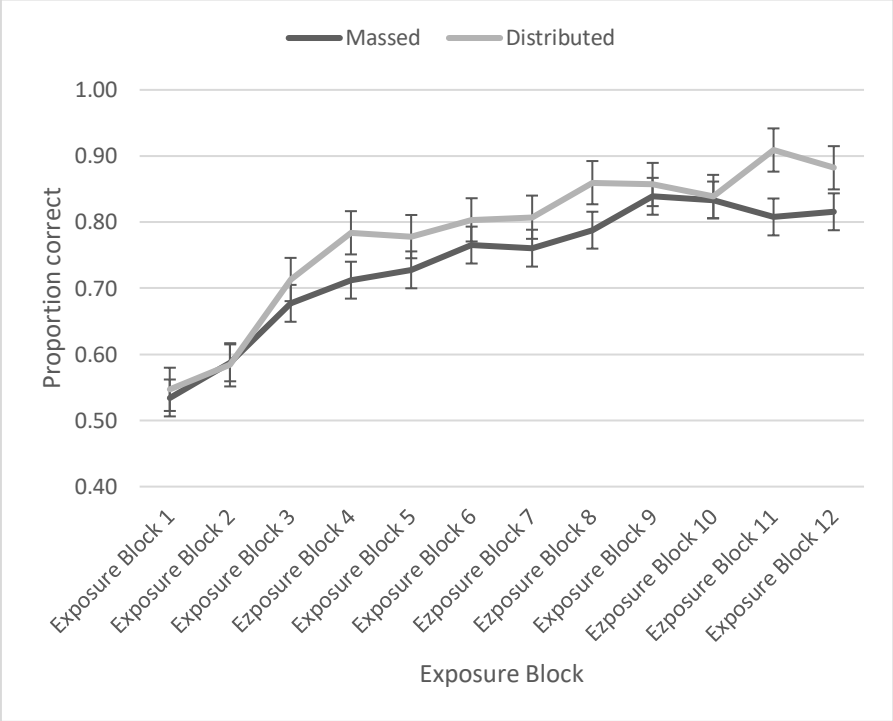
Each exposure block produced scores for the cross-situational learning task. See Table 8 for the descriptive statistics and Figure 8. Likewise, scores were produced for each of the five lexical categories over tests 1 to 5, with test 5 being the 24-hour delayed posttest. See Figure 9 and Table 9 for the descriptive statistics. In order to ascertain when learning had taken place during the exposure blocks, a one-sample t-test was conducted to compare the mean scores for each block to a chance score of .5. For both the massed and the distributed groups, participants performed significantly above chance from block two (massed (M = .57, SD = .15), 95% CI [.52 to .63], $t(31) = 2.64$, p = .006; distributed (M = .57, SD = .20), 95% CI [<.50 to .65], $t(63) = 2.06$, p = .024) onwards. Thus, 32 trials of exposure (without feedback) were enough to lead to above-chance performance in the cross-situational learning task for both groups. It should be noted that block 2 was before the first lag for the distributed group.

Table 8. Descriptive Statistics for Exposure Blocks

Exposure Block	Massed		Distributed	
	Mean	Std. Deviation	Mean	Std Deviation
Exposure Block 1	0.52	0.15	0.54	0.15
Exposure Block 2	0.57	0.15	0.57	0.20
Exposure Block 3	0.69	0.18	0.70	0.18
Exposure Block 4	0.73	0.21	0.79	0.17
Exposure Block 5	0.73	0.22	0.78	0.20
Exposure Block 6	0.78	0.21	0.80	0.18
Exposure Block 7	0.78	0.23	0.81	0.21
Exposure Block 8	0.79	0.26	0.85	0.18
Exposure Block 9	0.84	0.20	0.85	0.19
Exposure Block 10	0.83	0.24	0.84	0.21
Exposure Block 11	0.81	0.24	0.90	0.16
Exposure Block 12	0.83	0.23	0.87	0.17

In order to determine whether the distribution of practice affected learning during the training blocks, a repeated-measures ANOVA with a Greenhouse Geisser correction was carried out with blocks 4-12, excluding blocks 1-3 as they occurred before the first lag in the distributed group. Regarding block*group, no significant difference was found between the distributed and massed groups over blocks 4-12 [$F(4.703, 291.566) = 1.140$, $p = 34$].

Figure 8. Performance on the Cross-situational Exposure Blocks for Massed and Distributed Groups



Note. Both groups reached above chance by block 2. There was no significant difference between the performance of the two groups.

Table 9. Descriptive Statistics for Tests 1-5 for the Five Lexical Categories

	Massed		Distributed	
Test	Mean	Std. Deviation	Mean	Std. Deviation
Word order test 1	0.76	0.19	0.76	0.20
Word order test 2	0.81	0.23	0.84	0.16
Word order test 3	0.84	0.24	0.85	0.20
Word order test 4	0.84	0.24	0.86	0.20
Word order Delayed test 5	0.82	0.25	0.87	0.18
Noun test 1	0.51	0.21	0.54	0.15
Noun test 2	0.57	0.17	0.63	0.21
Noun test 3	0.64	0.27	0.72	0.21
Noun test 4	0.66	0.22	0.73	0.23
Noun delayed test 5	0.63	0.27	0.72	0.22
Verb test 1	0.71	0.25	0.68	0.26
Verb test 2	0.74	0.27	0.77	0.26
Verb test 3	0.78	0.26	0.74	0.30
Verb test 4	0.77	0.29	0.80	0.28
Verb delayed test 5	0.85	0.22	0.84	0.22
Adjective test 1	0.53	0.22	0.57	0.29
Adjective test 2	0.55	0.22	0.51	0.25
Adjective test 3	0.52	0.23	0.60	0.28
Adjective test 4	0.67	0.27	0.62	0.28
Adjective delayed test 5	0.58	0.25	0.59	0.22
Marker test 1	0.49	0.16	0.50	0.18
Marker test 2	0.53	0.18	0.52	0.13

Marker test 3	0.52	0.17	0.55	0.19
Marker test 4	0.49	0.18	0.52	0.25
Marker delayed test 5	0.56	0.15	0.52	0.17

Figure 9. Lexical Category and Word Order Tests

Note. Results of the lexical categories over tests 1 to 4 on day 1 and delayed test 5 on day 2, divided into massed and distributed groups.

4.6.2 24-hour Delayed Posttest

In order to determine if the massed and distributed exposure conditions affected results on the 24-hour delayed test, we first carried out one-sample t-tests to determine whether the mean score for the two groups (massed and distributed) reached above chance (at .5). See table 10. For word order, nouns and verbs, both massed and distributed groups were significantly above chance at test 5. Only the distributed group achieved an above chance score for adjective test 5, while only the massed group reached above chance for case markers test 5.

A MANOVA was conducted to examine the effects of massed and distributed groups on the combined dependent variables (syntax delayed test 5, verb delayed test 5, noun delayed test 5, adjective delayed test 5, and marker delayed test 5). The overall multivariate effect was statistically insignificant, Wilks' $\Lambda = 0.902$, $F(6, 58) = 1.259$, $p = .29$, $\eta^2 = .098$. Looking at individual delayed test scores, independent samples t-tests revealed no significant difference between massed and distributed groups for any of the lexical categories on delayed test 5. The results are summarised in Table 11

Table 10. Summary of One-Sample T-Tests of Delayed Test 5 for Massed and Distributed Groups

	Massed			Distributed		
Test	*T*	*df*	Sig. (2-tailed)	*T*	*df*	Sig. (2-tailed)
Word order test 5	7.21	31	<.001	11.824	31	<.001
Noun test 5	2.75	31	.010	5.695	31	<.001
Verb test 5	9.14	31	<.001	8.796	31	<.001
Adjective test 5	1.78	31	.084	2.370	31	.024
Case marker test 5	2.15	31	0.04	0.842	31	.41

Table 11. Summary of Independent-Samples T-Test for Delayed Test 5 with Massed and Distributed Exposure Groups as the Independent Variable

Test	*T*	*Df*	Sig. (2-tailed)
Word order test 5	-0.99	62	.33
Noun test 5	-1.51	62	.14
Verb test 5	0.21	62	.83
Adjective test 5	-0.20	62	.84
Case marker test 5	0.79	62	.43

In terms of the trajectory of change from the end of the learning phase on day 1, as measured by test 4, to the delayed test 5 the following day, for the massed group, there were significant increases in scores for case markers ($t(31) = -2.440$, $p = .02$, two-tailed) and approaching significance for verbs ($t(31) = -2.009$, $p = .53$, two-tailed). In order to determine if the trajectory of change from the end of day 1 to the delayed tests on day 2 was different between the massed and distributed groups, I carried out repeated measures ANOVAs. Table 12 shows the results. While none of the categories were significant, the two grammatical

categories of word order and case markers were approaching significance. For word order, the distributed group improved overnight while the massed group declined. For the case markers, both groups improved but the massed group had a greater improvement.

Table 12. Summary of repeated–measures ANOVA over Tests 4–5 showing effect for test block*Group (massed vs. distributed)

Effect	**df**	**F**	**p**
Word Order	**1. 62**	**1.83**	**.18**
Noun	**1. 62**	**.14**	**.71**
Verb	**1. 62**	**.63**	**.53**
Adjective	**1. 62**	**.83**	**.37**
Case Marker	**1. 62**	**.18**	**.19**

4.6.3 Individual Differences Measures

Descriptive statistics for the individual difference measures can be found in Table 13.

Table 13. Descriptive statistics for ID measures

Name of test	Massed			Distributed		
	Mean	Std. Deviation	Range	Mean	Std. Deviation	Range
SRT	43.2	33.4	146	65.1	34.8	170
MLAT-V	16.1	5.6	19	17.4	4.1	17
Aospan	36.5	19.7	75	39.6	19.2	72
CVMT	1.85	0.58	2.28	1.76	0.52	2.13
NRT	21.3	2.32	11	21.5	3.0	12

4.6.3.1 Individual Differences on Day 1, tests 1-4. First, in order to investigate the prediction that procedural memory, as measured by SRT, would predict success during the exposure training, a step-wise regression was calculated with the overall average score for exposure blocks 1-12 as the dependent variable and the five individual difference measures (MLAT-V, SRT, CVMT, Aospan, NRT) as the independent variables. For the massed group, success was predicted by CVMT, $R^2 = .129$, $F(1, 31) = 4.433$, $p = .044$, while for the distributed group SRT predicted success, $R^2 = .176$, $F(1, 30) = 6.198$, $p = .019$.

Next, I carried out exploratory step-wise regressions for tests 1-4 with individual difference measures as independent variables. I decided to include each lexical category and word order test results for each test rather than reducing the number of dimensions through factor analysis in order to give a more fine-grained picture of the role the individual differences play for both massed and distributed groups. Figure 10 shows the predicting individual difference measures from the regressions and the full regression table can be found in appendix H.

Figure 10. Predicting Individual Difference Measures

	Word Order		Verbs		Nouns		Adjectives		Markers	
	Massed	Distributed	Massed	Distributed	Massed	Distributed	Massed	Distributed	Massed	Distributed
Test 1	MLAT-V									
Test 2	MLAT-V					MLAT-V	Aospan	SRT		SRT
Test 3										Aospan
Test 4	MLAT-V					SRT	SRT & NRT	SRT		NRT
Test 5	MLAT-V					SRT		MLAT-V		MLAT-V
Test 1-4 average	MLAT-V	CVMT		-CVMT -NRT			NRT	CVMT SRT		

	Massed	Distributed
Overall CSL score	CVMT	SRT

Note. Predicting individual difference measures for tests 1-5, the average of test 1-4 and the average score for the cross-situational learning exposure task. The full regression table can be found in Appendix H. MLAT-V = part 5 of the Modern Language Aptitude Test (verbal declarative memory); CVMT = Continuous Visual Memory Test (Visual declarative memory); SRT = Serial Reaction Time (procedural memory); NRT = Non-Word Repetition Task (phonological short-term memory); Aospan = Automated Operation Span Task (working memory capacity).

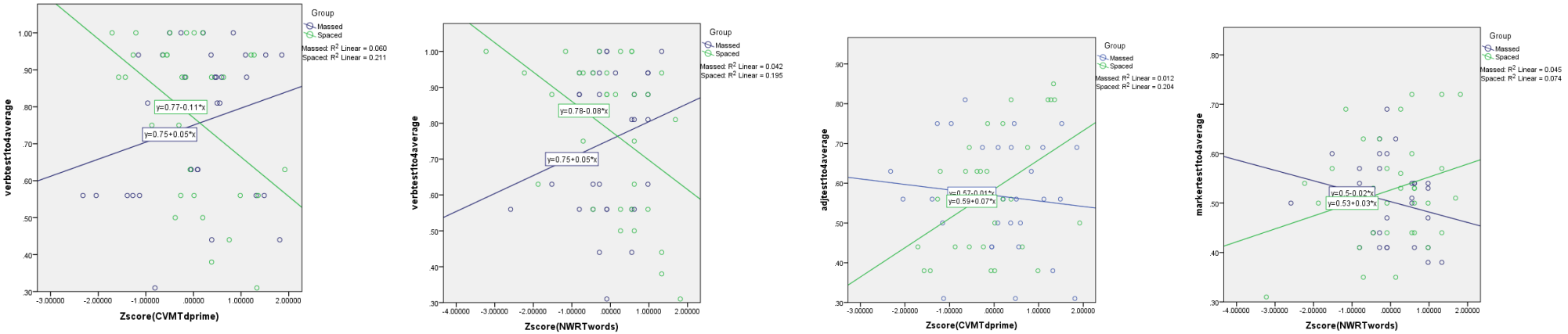
I then calculated interactions between group (massed or distributed) and individual difference measures for tests 1-4. Four interaction effects were found from the regression analyses. Table 14. and Figure 11 show the results. Firstly, group and CVMT interacted for the average of verb tests 1-4. Likewise, group and NRT scores also interacted for the average of verb tests 1-4. That is, those with better CVMT scores (visual declarative memory), and also those with better NRT scores (phonological short-term memory), did better on the average of verb tests 1-4 if they were in the massed group but worse if they were in the distributed group.

An opposite interaction was found for the average of adjective tests 1-4, for which there was an interaction between CVMT scores and group. That is, those with better CVMT scores did better on the average of adjectives tests 1-4 if they were in the distributed group but worse if they were in the massed group. For the average of case marker tests 1-4, there was an interaction between NRT and group. Those with better NRT scores did better on the average of case markers tests 1-4 if they were in the distributed group but worse if they were in the massed group.

Table 14. Stepwise Linear regressions showing all interactions that were found between ID measures and group (massed or distributed) for the average of test 1-4 for each lexical category and word order.

Test	Predicting variables	*B*	*SEb*	*B*
Verb tests 1-4 average	(Constant)	0.767	0.023	
	NRT x Group	0.062	0.024	0.304*
	CVMT x Group	0.057	0.024	0.278*
Adjective tests 1-4 average	(Constant)	0.576	0.016	
	SRT	0.047	0.016	0.326*
	CVMT x Group	-0.046	0.016	-0.321**
	NRT	0.033	0.016	0.228*
Case marker test 1-4 average	(Constant)	0.515	0.012	
	NRT x Group	-0.025	0.012	-0.252*

Figure 11. Interactions between Individual Difference Measures and Group



Note. Interactions between individual difference measures and group (massed and distributed) for the average score of tests 1-4. Left: the interaction between CVMT and group for the average score of verb tests 1-4. Middle left: the interaction between NRT and group for the average score of verb tests 1-4. Middle right: The interaction between CVMT and group for the average score of adjective tests 1-4. Far rright: the interaction between NRT and the average score of case markers 1-4.

4.6.3.2 Individual Differences and the Delayed Posttests. In order to determine whether individual difference measures predicted success on the final delayed-test 5 for the four lexical categories and for word order, I carried out step-wise linear multiple regressions for both massed and distributed groups, with the score for test 5 for each category as the dependent variable and the five individual difference measures (SRT, MLAT-V, CVMT, Aospan, NRT) as independent variables. As can be seen from Table 15, there were predictor variables for three lexical categories in the distributed group: nouns (SRT), adjectives (MLAT-V) and markers (MLAT-V) and for word order (MLAT-V) in the massed group.

Table 15. Stepwise Linear Regressions for delayed test 5 showing predicting ID measures for lexical categories and word order for massed and distributed groups

Component		Massed			Distributed		
		B	*SEb*	*B*	*B*	*SEb*	*B*
Word Order test 5	Constant	0.83	0.042				
	MLAT-V	0.083	0.038	0.37*			
Nouns test 5	Constant				0.69	0.039	
	SRT				0.10	0.039	0.44*
Adjectives Test 5	Constant				0.59	0.036	
	MLAT-V				0.12	0.043	0.47*
Case markers test 5	Constant				0.513	0.028	
	MLAT-V				0.090	0.033	0.45*

Note. *= p < .05; ** = p< .01; *** = p< .001

4.7 Discussion

4.7.1 Massed vs. Distributed Practice

The principal aim of this experiment was to investigate whether the durability of language learning over 24-hours could be influenced by manipulating the distribution of language exposure. The results were not clear-cut. On the one hand, my prediction that results on the 24-hour delayed tests for the different lexical categories would be significantly better for the distributed group than the massed group was not supported. Several factors could account for this. Firstly, the cross-situational learning conditions and the artificial language itself may have been too complex to benefit from the distributed practice effect (Donovan & Radosevich, 1999). A meta-analysis performed by Donovan and Radosevich (1999) found a much smaller effect size for more complex tasks (d= .07) than simple tasks (d= .97). However, as Rogers (2017) points out, it may not be easy to generalise these findings to SLA due to their operationalisation of complexity. This involved clustering studies into four groups according to the number of (physical and mental) operations required to perform a task (e.g., air traffic control for a complex task; climbing a ladder for a simple task). Only the most complex tasks, according to their criteria, involved a high level of mental complexity, and this group had almost the same effect size as the next highest level that did not include mental complexity (d= .11). This raises the question of to what extent their meta-analysis highlights that L2 complexity affects the size of the distributed practice effect. Another meta-analysis by Janiszewski et al. (2003) found that distributing compared to massing presentations of structurally complex stimuli resulted in a similar average effect size (M=.330) to semantically simple stimuli (M=.325). Semantically complex stimuli, on the other hand, had a significantly larger average effect size (M=.586) than semantically simple stimuli. While the example Janiszewski et al. (2003) gave for structurally

complex stimuli was a simple SVO sentence (The cat is on the red brick wall) and therefore arguably not as complex as the SOV/OSV plus nonce words contained in the artificial language in this study, their findings suggest that structural complexity may not dampen the distributed practice effect but enhance it. Indeed, they found that the average effect size for nonsense words (M=.454) was greater than for real words (M=.330). From a desirable difficulty perspective, it is possible that the cross-situational learning task, which involves a combination of implicitly keeping track of statistical probabilities and more explicit hypothesis testing, may have been too demanding as unsuccessful inferences made during the training phase may have interfered with the correct forms (Suzuki et al., 2019).

Another possible factor for why there was no statistically significant difference between distributed and massed groups on delayed test scores is that incidental learning conditions have less of a distributed practice effect than intentional learning conditions (Janiszewski et al., 2003). Study 1 involved a cross-situational learning paradigm, wherein participants not only were not told that there were rules governing syntax and word categories or explicitly taught any of the word referents, but they were also not given any feedback on choices or told there would be a test the next day. This experimental design mimics immersion conditions that contain little if any explicit instruction, and the learning conditions are arguably more incidental than some previous distributed practice SLA studies that used an incidental learning condition (e.g., Rogers, 2015). In Rogers (2015) study, university students were presented with sentences one at a time on the classroom board, which may have encouraged a more conscious attention to form, despite the instructions of the task directing the participants towards comprehension. However, the meta-analysis by Janiszewski et al., 2003) did still find a spacing effect for incidental learning

conditions (average effect size of .236), as have a number of other studies (Greene, 1989; Rogers, 2015; Verkoeijen, et al., 2005).

Finally, it is possible that a different ISI/RI ratio may have resulted in a significant difference between the massed and the distributed groups. While the ratio chosen for this study (1.3%) was based on findings by Kornmeier et al. (2014) that for a 24-hour RI a 20-minute ISI was optimal, other researchers have postulated optimal ratios of 20-40% ISI/RI for shorter ISIs (7 days) reducing to 5-10% for very long (1 year) ISIs (Cepeda, et al., 2008). These latter ratio guidelines were followed for the distributed groups in the recent studies on the distributed practice effect and L2 grammar by Bird (2010; ISI/RI ratio of 25% for the distributed condition), Rogers (2015; 17%), Suzuki and DeKeyser, (2017a; 25%). For study 1, an ISI/RI ratio of 20-40% would mean a lag of 4.8 – 9.6 hours. It should be noted, however, that Cepeda et al. (2008) used a paired-associates paradigm and that optimal spacing for grammatical language items presented under incidental learning conditions may not be the same. Firstly, there is evidence from distributed practice studies that incidental learning conditions require a narrower optimal lag for a given retention interval than intentional learning conditions (see meta-analysis by Janiszewski et al., 2003). Secondly, Kim et al. (2013) proposed a skill retention theory in which procedural knowledge does not benefit from distributed practice as procedural memory traces decline at a much slower rate than declarative memory. This is an area I will return to in study 2. It would be useful for a future study to return to the cross-situational learning paradigm and investigate whether an ISI/RI ratio of 20-40% produces a distributed practice effect.

Despite a lack of an overall significant difference between the massed and the distributed groups for the lexical and word order tests, two small differences between the conditions suggest that manipulating the distribution of exposure may play at least some role in the speed and

effectiveness of acquisition. Only the distributed group reached above chance on the adjective delayed test 5 and only the massed group were above chance for the case marker delayed test 5. In addition, for the massed group, there was a significant increase between the end of learning on day 1 and the delayed posttest on day 2, but this was not the case for the distributed group. Suzuki and DeKeyser (2017) found that for Japanese morphosyntax the massed condition had better reaction times for oral production, and they suggested that this was because the reaction time test was tapping into procedural rather than declarative knowledge. It is possible that the grammatical, incidentally presented case markers benefitted more from unbroken massed exposure than the more idiosyncratic, lexical referents for the adjectives. It would be interesting to repeat the study with a wider ISI/RI ratio, and with another block of practice, to investigate whether these small differences enlarge with more time and a more pronounced ISI/RI ratio.

4.7.2 Individual Differences x Massed and Distributed Practice

The second aim of this study was to investigate the role that five cognitive individual difference measures (verbal and visual declarative memory, procedural memory, working memory capacity and phonological short-term memory) played in the learning of the artificial language under cross-situational learning conditions and the way that they interacted with massed and distributed learning conditions.

No regular pattern emerged regarding which individual difference measures predicted success on the different lexical categories at test 5. The hypothesis that working memory capacity and phonological short-term memory would predict success on delayed tests under massed conditions was not supported. Previous research (Suzuki & DeKeyser, 2017b; Verkoeijen and Bouwmeester (2008) suggests that limited working memory capacity may cause problems encoding to-be-learned items when under massed conditions. One possible reason for

the lack of interaction between working memory and massed groups may be that the speed of presentation was not sufficiently demanding to force a reliance on working memory capacity (DeKeyser, 2012; Verkoeijen and Bouwmeester, 2008). However, post-hoc analysis shows that CVMT, the measure of visual declarative memory, and Aospan, the measure of working memory capacity, correlated, ($r = .378$, $p = .002$) and interrelated when a principal components analysis was performed on all five individual difference measures, so it is possible that there is some overlap between the measures, which would therefore suggest that the finding that CVMT predicts success on the word order delayed posttest 5 on day 2 in the massed group, and that CVMT may capture some elements of visual working memory, if not the ability to manipulate visual memories. It is also interesting to note that during exposure training (on an average of tests 1-4), NRT, a measure of phonological short-term memory was a predicting variable for adjectives.

Scores for the 24-hour delayed tests 5 in the distributed group were more affected by the individual difference measures than in the massed group. In the distributed group, the three lexical categories that constituted the noun phrase were predicted by individual difference measures: nouns by SRT, adjectives and case markers by MLAT-V. In contrast, in the massed group only word order had a predictor variable: MLAT-V. Is it possible that distributing the presentation of language increases the role of (some) individual differences?

DeKeyser (2012) provides a general rule of thumb for aptitude x treatment interactions: the more an instructional treatment forces a learner to rely on a particular attribute, the more individual differences will play a role. That is, specific instructional treatments can allow or encourage particular aptitudes to become vital where alternative treatments might not. An example of this comes from the study by Erlam (2005), who investigated potential interactions

between three individual cognitive differences (language analytic ability, working memory and phonemic coding ability) and three instructional methodologies (deductive instruction, in which grammar rules were explicitly taught to the participants; inductive instruction, in which participants had to search for rules themselves from exemplars; and structured input instruction, in which participants were given receptive practice of the grammar rules) for 92 New Zealand high school students learning the direct object pronoun in French. She found that there were many fewer interactions (as measured by correlations) between the individual difference measures and test results in the deductive instruction condition than the other two conditions. She concluded that deductive instruction with opportunities to produce the target structure neutralised individual differences, while the other two instructional methods, with less scaffolding, forced learners to rely more on their cognitive abilities and thus there were more individual differences found. Returning to the results of this study, it is possible that for this complex procedural task, remembering the (mostly partially known for most of the study) adjectives and case markers in memory over the 20-minute lag forced a reliance on visual declarative memory, whereas this individual difference measure was not called upon as much in the massed condition.

Regarding the hypothesis that procedural memory would predict success during the exposure training in the distributed group but not the massed group, this was supported for the performance over the exposure training trials, albeit not for the individual lexical category test trials. A number of studies have shown an advantage for distributing practice for more procedural tasks during learning rather than on delayed posttests (Lee & Genovese, 1988; Kwon et al., 2015; Mackey et al., 2002; Moulton et al., 2006), possibly due to offline consolidation (Fahl et al., 2023; Robertson et al., 2004; Schönauer et al., 2014). Differences in procedural

memory ability, therefore, may influence the amount of offline consolidation that occurs. A future study that investigates the interaction between individual differences in procedural memory and offline consolidation of procedural language tasks could be beneficial in shedding light on the mechanisms underlying the distributed practice effect for more procedural knowledge and tasks.

Continuing the exploratory analysis, I investigated interactions between the five individual difference measures and group (massed /distributed) for the average of tests 1-4 for each lexical category. Several interesting interactions emerged. CVMT (visual declarative memory) and NRT (phonological short-term memory) interacted with group for the average of verb tests 1-4. Those with better visual declarative memory and phonological short-term memory did better in the massed groups but worse in the distributed groups. In contrast, there were two other interactions that showed a benefit for distributed exposure and a detriment for massed exposure: between CVMT (visual declarative memory) and the average of adjective tests 1-4; and between NRT (phonological short-term memory) and the average of case marker tests 1-4. That is, those with better visual declarative memory did better on average adjective tests 1-4 in the distributed group but worse in the massed group. And those with better phonological short-term memory did better on the average of case marker tests 1-4 in the distributed group but worse in the massed group.

Why might stronger declarative memory and phonological short-term memory result in worse scores on the verb tests but better scores on the adjective tests when in the distributed group but the opposite when in the massed group? One possible explanation is that spacing either allowed or caused those with better visual declarative memory and better phonological short-term memory to shift their focus away from verbs and their sentence-final position and

notice the noun phrase in general and adjectives and case markers respectively in particular. With attention thus shifted away from verbs and towards the learning of the noun phrase, learning of the verbs may have suffered in comparison to those with worse visual declarative memory and phonological short-term memory. Instead, those with worse visual declarative and phonological short-term memories, who may have been less able to shift their focus to other lexical categories, particularly after the first break, may have instead continued to focus on learning the verbs all the way through the exposure blocks, thus giving them an advantage with this lexical category.

Circumstantial supporting evidence for this explanation can be found from drilling down into the predictors of success for individual lexical tests. Follow-up regressions on individual tests revealed that success in word order test 1 was predicted by CVMT, a measure of declarative memory. The break then possibly shifted those participants' attention to the noun phrase, as noun test 2 is predicted by CVMT only for the distributed group. Noun test 2 then predicted success with noun test 4.

The question remains as to why a 20-minute lag would help those with better visual declarative memory and phonological short-term memory shift attention from verbs and word order to the noun phrase. Possible explanations include one that involves maintenance and forgetting of memories; and another involving the initial encoding of those memories and/or proceduralisation. Regarding maintenance and forgetting, strength of visual declarative memory (and phonological short-term memory) may have directly influenced the rate of forgetting during the lag. After an undesirably difficult 20-minute lag, those with poorer visual declarative memory (and phonological short-term memory) may have forgotten too much of the verb and word-order form-meaning connections to be able to either consciously or subconsciously shift

attention to the noun phrase. Whereas, for those with better visual declarative and phonological short-term memories, a more desirably difficult 20-minute lag would have resulted in enough of the verbs to be remembered to then shift their attention to other aspects of the artificial language.

Another, potentially complementary, explanation regards the initial encoding before the lag. Given that word order test 1 was predicted by scores of CVMT (visual declarative memory), stronger visual declarative and phonological short-term memory may have resulted in learning beginning to reach mastery (Mettler et al., 2020), beginning to proceduralise the knowledge (Kim et al., 2013) or even just reaching explicit cognition (Cleeremans, 2007, 2011) before the break. This qualitative change may have then resulted in a lower rate of forgetting over the 20-minute lag than for those who did not reach a level of mastery before the break, or in the case of reaching explicit cognition, reaching a learning phase which benefits from distribution of exposure. Next, in chapter 3, we suggested that shifting attention may act in a similar way to changing the learning conditions from incidental to intentional in that it might increase the reliance on declarative memory (Ullman and Lovelett, 2018). If this were the case, strong declarative memory would not be causing a shift in attention in the distributed group, but rather participants' reliance on declarative memory would be the by-product of the shift in attention, which may be caused by other factors related to the 20-minute lag. Finally, one intriguing possibility is that if CVMT does indeed capture some aspects of visual working memory, then shifting attention may not be the result of declarative memory differences but visual working memory. Kapa et al., (2017), in a study involving pre-schoolers with specific language impairment, found links between visual working memory and sustained visual attention (see also Smolak et al., 2020). Further research in this area is needed.

4.8 Limitations

It needs to be remembered that much of the analysis of individual differences was exploratory. With few specific hypotheses set and a number of different regressions performed on different lexical categories at five different testing points, the chances of type 1 errors occurring increase. Therefore, these results should be considered with caution, but provide a potentially interesting starting point for larger-scale studies. Further research should investigate the role cognitive individual differences play in shifting attention or developing insights during cross-situational learning, particularly as regards how they interact with spacing conditions.

4.9 Summary of Chapter 4

In this chapter, I have reanalysed the data from study 1 to investigate whether distributing exposure to an artificial language under cross-situational learning conditions affects the durability of learning after 24 hours. The hypothesis that distributed schedules would result in better test results after 24 hours than massed schedules was not met. However, there is some evidence that differences in learning took place under massed and distributed schedules. These differences are that individual memory differences appeared to play a larger role under the distributed practice schedule. In addition, procedural memory predicted success of learning during exposure training blocks under the distributed practice schedule. The hypothesis that working memory capacity would predict success of lexical categories under massed conditions was not met. However, the hypothesis that declarative memory would predict success for lexical categories under distributed conditions was partially met. Markers and adjectives were predicted by declarative memory but verbs and nouns were not. Finally, interactions between schedule and individual difference measures suggest that the lag may cause or allow a shift in attention, which then influences the extent to which particular lexical categories are learned.

The next chapter reports study 2, in which I built on the findings of study 1 to investigate several factors (intentional vs. incidental learning conditions; items that require a generalisation of the rules vs. items that appeared in the exposure blocks; and declarative memory) that may influence the optimal lag when learning form-meaning connections. Study 2 used more educationally relevant time periods (lags in days and a 35-day retention interval) and included more lag groups (five) than study 1.

Chapter 5: The Optimal Lag for Intentional and Incidental Language Learning

5.1 Introduction

In chapter 3 and 4, I described and analysed study 1, which investigated firstly in chapter 3 whether an artificial language with nouns, verbs, adjectives, case markers and a verb-final syntax could be learned via a cross-situational learning paradigm, what the order of acquisition was, and whether learning was durable after 24 hours. In chapter 4, I then reanalysed the data to investigate whether distributing the exposure of the artificial language under cross-situational learning conditions resulted in better learning and retention over 24 hours than massing. Furthermore, I investigated whether five individual memory differences (verbal and visual declarative memory, procedural memory, working memory capacity and phonological short-term memory) differently affected the learning of the artificial language under massed and distributed schedules. Results showed that not only can the artificial language be learned under cross-situational learning conditions, with verbs and basic word order learned first, followed by nouns, adjectives and finally case markers, but learning was also durable over 24 hours. In terms of the distributed practice effect, results showed that distributed schedules did not result in significantly better test results after 24 hours than massed schedules, although there was some evidence that there were differences in learning taking place under massed and distributed schedules. Distributed practice schedules appeared to encourage or force a reliance on individual memory differences, and this may cause or allow a shift in attention. Declarative memory predicted success of some lexical categories in the distributed group but not all. Procedural memory predicted success of learning during the exposure phase in the distributed group.

In the current chapter, I describe study 2, in which I build on the results of study 1 by investigating several factors that may influence the optimal lag for a given educationally relevant

RI. First, I will outline some of the key literature regarding the optimal lag of L2 grammar, then the methodology, results and discussion follow.

5.2 Review of Relevant Literature

5.2.1 Background to the Study

Practice makes perfect, but when and how often should you practise? This issue, the optimal temporal gap between practice sessions to enhance learning and memory retention, is of critical interest to language learners, teachers, syllabus designers, materials developers as well as online learning app designers. The current study investigated whether one factor (intentional vs. incidental learning conditions) influences the optimal spacing between practice sessions when learning an aspect of L2 grammar.

Over the past couple of decades, a growing number of studies have investigated the optimal spacing of facts (Cepeda et al., 2008), rereading expository texts (Verkoeijen et al., 2008), word pairs (Küpper-Tetzel & Erdfelder, 2012) and L2 vocabulary (Küpper-Tetzel et al., 2014; Nakata, 2015). In many domains of learning, the intersession interval (ISI) between practice sessions has a strong effect on memory retention, with spaced practice better than massed (i.e., with no time between study sessions) (see Delaney et al., 2010, for an overview). Moreover, the width of the lag appears to have a profound effect on learning. A meta-analysis by Cepeda et al. (2006) showed that the ISI displayed non-monotonic effects – an inverted u-shape – over time (see also Crowder, 1976/2014; Glenberg & Lehmann, 1980). That is, widening ISI results in better scores, but ISIs wider than optimal then produce worse results. Their study also suggested that the optimal ISI depends on how long you want to remember for.

In an influential follow-up study, Cepeda et al. (2008) investigated the learning and retention of fun trivia facts by 1350 participants over two sessions and then tested them on a

delayed test. They had 24 conditions with six lag groups with varying ISIs from 0 to 105 days and four retention intervals (RI) from 7 to 350 days. They found that the optimal ISI depends on the RI. They found that for short RIs, the optimal ISI was around 20-40% of the RI and for the longest RIs, the ratio fell to around 5-10%. For a 35-day RI, used in the current study, the optimal ISI was the 7-day ISI (further estimated as a 6-day optimal ISI using interpolated cubic splines), which equates to a 19% ISI/RI ratio. Figure 3 in section 2.3.1 summarises the results. Rohrer and Pashler (2007), drawing on the data from Cepeda et al. (2008), suggested a rule of thumb of 10-30% range of optimal lag. This 10-30% optimal range has since been used in much subsequent research, including SLA studies.

5.2.2 The Distributed Practice Effect and L2 Grammar

Since the seminal study by Cepeda et al. (2008), there have been a handful of studies that have investigated distributed practice for L2 grammar learning as opposed to learning vocabulary or facts using Rohrer and Pashler's optimal ISI/RI ratio of 10-30% (Bird, 2010; Kasprowicz et al., 2019; Miles, 2014; Rogers, 2015, Suzuki & DeKeyser, 2017a; Suzuki, 2017). The majority of these studies were designed with two ISIs and two RIs, with one group positioned inside the optimal ISI/RI ratio (e.g., 25%) and one group outside the optimal ratio (e.g.,100%) for each RI. However, these L2 grammar studies have produced varying results.

Classroom-based studies by Bird (2010), Miles (2014) and Rogers (2015) showed distributed practice effects for L2 grammar using the optimal spacing ratio laid out by Cepeda et al. (2008). Bird (2010) found that distributing the practice and error correction of verb tenses with an ISI of 14 days produced significantly improved results on a delayed grammaticality judgment test of 60 days, compared to a shorter non-optimal ISI of 3 days. Miles (2014) found that distributing the practice of the word order of English adverbs at an expanding ISI of 7 to 28

days resulted in greater gains than massing practice for an editing test but not for a translation test after 35 days. Rogers (2015) found a distributed practice effect for the incidental presentation of the syntax of five English cleft sentence structures with an ISI of 7 days when tested after 42 days compared to a shorter ISI of 2.25 days.

In contrast to these three studies, several other studies have found no benefit for distributed practice for L2 grammar at ISI/RI ratios of 10-30%. Suzuki and DeKeyser (2017a) found no difference in accuracy measures of Japanese morphosyntax between an ISI of 1 day and an ISI of 7 days when tested after both 7 days (ISI/RI ratio of 14% for the 1-day ISI and 100% for the 7-day ISI) and 28 days (3% for the 1-day ISI and 25% for the 7-day ISI). Indeed, reaction time measures were actually better for the shorter ISI at the 18-day RI posttest. In a follow-up study using a miniature language, Suzuki (2017) found that accuracy was significantly better on a 3.5-day ISI than a 7-day ISI for RIs of both 7 and 28 days, although a re-analysis to include the potentially confounding posttest as another study session showed that the shorter, now 4.25-day ISI and 21-day RI with an ISI/RI ratio of 20%, were not far off the optimal ratio of 17% for a 35-day RI from Cepeda et al. (2008). A study by Kasprowicz et al. (2019) looking at the learning of French verb inflections by eight-year-old children also found no difference between a 2.5-day ISI group (6.0% ISI/RI) and a 7-day ISI group (17% ISI/RI) on two 42-day RI delayed tests.

A recent meta-analysis of the effects of distributed practice on second language learning by Kim and Webb (2022) included 98 effect sizes from 48 experiments involving both grammar and vocabulary. They found that distributed practice is superior to massed practice ($g = 0.58$, a medium-to-large effect size). They also found that longer lags are superior to shorter lags when tested on a delayed posttest ($g = 0.40$, a medium effect size). They found that type of activity

moderated the distributed practice effect, with longer lags better than shorter lags for L2 grammar (g= 0.56). However, these results should be taken with caution. Longer and shorter lags were only coded with respect to the comparison groups within each study. Therefore, a 2-day lag may be considered a shorter lag in one study but a longer lag in another.

One possible reason for the mixed results in these L2 grammar studies is that there may be other factors that influence the optimal ISI besides its relationship to the RI. Suzuki and DeKeyser (2017a) suggested that the increased complexity of oral productive tasks in their two studies compared to the receptive tasks in Bird's (2010), and Rogers' (2015) studies may have been more susceptible to skill loss at the longer (7-day) ISI groups. Suzuki and DeKeyser (2017) found that results on the delayed test at an RI of 28 days for accuracy were starting to converge and they speculated that a wider RI, and therefore a smaller ISI to RI ratio, might have seen the 7-day ISI outperform the 1-day ISI group. Another aspect of Suzuki's (2017) study investigated the role that item complexity has on the optimal ratio of ISI to RI, but they did not find any significant effect.

Situating their theory within the desirable difficulty framework (Bjork, 2018; Schmidt & Bjork, 1992; see also the reminding account by Benjamin & Tullis, 2010), which states that practice which is optimally difficult results in greater long-term gains, and the cognitive difficulty framework (Housen & Simoens, 2016), which distinguishes between linguistic difficulty, learner-related difficulty, and context-related difficulty, Suzuki et al. (2019) postulated that several other factors may also affect the optimal ISI for a given RI. These included whether knowledge is receptive or productive, the number of training sessions and the type of linguistic knowledge. In addition to these factors, a search of the spacing literature suggests additional factors which may also influence the optimal ISI / RI ratio. The following section highlights

three underexplored factors that are the focus of the current study. First, intentional and incidental learning instruction conditions are central to second language acquisition (SLA) research on grammar learning (Hulstijn, 2015, Rebuschat, 2015, Williams, 2005). However, it is chronically under-investigated in terms of the effect on the spacing of learning material. Second, the generalisation of patterns and their application to new contexts is critical to determining the effectiveness of grammar learning. The ability to generalise is the essence of flexible grammar use (Bod, 2009; Goldberg, 2006; Romberg & Saffran, 2010). Third, interest in individual memory differences, and particularly declarative, has increased significantly over recent years (Ullman, 2001; Walker et al., 2020).

5.2.3 Intentional and Incidental Learning Instruction Conditions

The effect that intentional and incidental learning instruction for to-be-learned items has on distributed practice has been an area of broad interest in cognitive psychology (Greene, 1990; Janiszewski et al., 2003; Toppino & Bloom, 2002; Verkoeijen et al., 2005). In the field of SLA, learning under intentional or incidental conditions is a matter of central importance, given the ongoing interest in both classroom-based teaching and acquiring language under more naturalistic, immersion conditions (see Hulstijn, 2012 for a short overview). When operationalised, under intentional learning conditions, the participant is told about what to learn, that they are to be tested and given any associated rules. Under incidental learning conditions, participants do not know they are to be tested and attention during exposure is misdirected away from remembering the word or learning the rule and towards another aspect of the input, for example the meaning.

A number of experiments have found a spacing effect for incidental learning conditions (Challis, 1993; Glenberg & Lehman, 1980; Greene, 1990; Greene & Stillwell, 1995; Jensen &

Freund, 1981; Shaughnessy, 1976). Over three experiments, Greene (1990) found a spacing effect on three tests of implicit memory: spelling of homophonic words, word-fragment completion and perceptual identification. It should be noted, however, that each of these experiments involved the learning of word lists in which the spaced condition involved a lag of only a handful of words (e.g., 5-8 words in experiment 1, that is, around 30 seconds), and the posttest was immediate. In none of these incidental spacing studies were educationally relevant lags and retention intervals used.

Incidental learning conditions appear to show a weaker distributed practice effect and require a narrower optimal lag than intentional learning conditions. A meta-analysis by Janiszewski et al. (2003) concluded that intentional learning conditions resulted in a stronger distributed practice effect than incidental learning conditions. However, the meta-analysis did not control for or investigate interactions with different length ISIs or ISI/RI ratios, leaving open the possibility that the optimal ISI is shorter under incidental learning conditions. In a subsequent study to investigate the relationship between ISI and incidental and intentional learning conditions for word learning, Verkoeijen et al. (2005) found that on a recall test, the optimal ISI was narrower for the incidental condition compared to the intentional condition in addition to having smaller overall distributed practice effects at their respective optimal ISIs. However, several aspects of this study mean that caution should be exercised before generalising to L2 grammar learning: similar to Greene (1990), the study involved a declarative word remembering task rather than a procedural grammar learning task; the ISIs were in seconds rather than days; and the test was carried out immediately following the exposure phase rather than after a delay.

Several studies have investigated whether distributed practice of L2 language learning under incidental learning conditions is beneficial, including contextual vocabulary learning (Elgort & Warren, 2014; Macis et al., 2021; Serrano & Huang, 2018; Webb & Chang, 2015) and L2 grammar learning (Rogers, 2015). Only Macis et al. (2021) compared both intentional and incidental learning conditions.

Macis et al. (2021) investigated the learning of L2 collocations under incidental and intentional learning conditions across two experiments. Participants were given one text (a short story in the incidental experiment and lines of concordance in the intentional experiment) to read a week for five weeks that either contained one collocation embedded five times in the text or lines of concordance (massed group) or 25 collocations embedded once in the text or lines of concordance (spaced group). In the incidental experiment, participants' completed comprehension questions (not involving the collocations) about the text. In the intentional learning experiment, participants were told to deliberately study the collocations. In both experiments, a cued form recall task was administered three weeks after the last treatment session (ISI/RI ratio of 33% for the distributed group). Results revealed that spacing had a large effect in the intentional learning experiment and a small effect in the incidental learning experiment. However, in the incidental experiment, the spaced condition was not as effective as the massed. They hypothesised that the difference in test results may lie in the amount of noticing that took place. Massing the target collocations within one text, they postulated, made them more salient and more likely to be noticed than in the spaced incidental condition (Schmidt, 1990, 2010). For the deliberate practice condition, as participants were asked to focus on the underlined target collocations in short concordance lines, there was a much greater chance of noticing, even in spaced conditions. These results suggest that massed conditions, or at least

shorter lags, may be more effective under incidental learning of L2 collocations. However, this study involved L2 vocabulary and did not look at the abstraction and transfer of rules that can be found in L2 grammar.

5.2.4 Generalisation

Many spacing studies, including the Cepeda et al. (2008) study that inspired the rule of thumb for ISI/RI ratios of 10-30% (Rohrer & Pashler, 2007), involved the learning and retention of declarative facts, but another important feature of language learning is the ability to generalise from rules. Does spacing out the presentation and practice of to-be-learned items increase the likelihood that generalised rules can be learned and then transferred to other learning situations? And if this is the case, is the optimal ISI narrower than for items that just need to be remembered? Research from various domains suggests that distributing the presentation and practice of more complex tasks improves the abstraction of learning and transfer to different learning contexts and tasks (Rohrer & Taylor, 2006; Vlach et al., 2012; Vlach et al., 2008; see Carpenter et al., 2012, for a review). Vlach et al. (2008) used a category-induction word-learning task in which made-up objects that differed in several aspects but kept the same shape were presented to three-year-old children in either massed or spaced conditions. Those in the spaced condition were better able to abstract the core meaning and identify a new version of that object on a multiple-choice posttest. A memory task in which the children were presented with the same object and just had to remember also found that spaced practice was better than massed but produced significantly better results than the category-induction task. Vlach et al. (2012) proposed a forgetting-as-abstraction theory to explain why distributing the presentation of exemplars promotes abstraction of rules. Upon encountering a second exemplar after a period of

time, generalised aspects of the to-be-learned items are strengthened, while non-generalised aspects continue to be forgotten. Therefore, over time, rules are abstracted.

In each of the L2 grammar studies mentioned above (Bird, 2010; Miles, 2014; Rogers, 2015; Suzuki & DeKeyser, 2017a), rules that had been given to participants were required to be applied to new exemplars in the testing phases. However, only Rogers (2015) involved incidental learning conditions in which the learners had to abstract the rules for themselves. No study, to our knowledge, has contrasted items previously trained (i.e., a test of memory) and generalisation of rules to new exemplars as within-subject variables to test optimal ISI/RI ratios.

5.2.5 Declarative Memory

Effects of learning conditions have so far been described, but there are also learner characteristics that are likely influential in the optimal lag for any given retention interval. Declarative memory is a long-term memory system that has been implicated in the learning of idiosyncratic items, such as L2 vocabulary (e.g., Ruiz et al., 2021), but also grammar at the early stages of acquisition (Hamrick, 2015). It is often measured using part 5, the paired-associates test from the Modern Language Aptitude Test (MLAT-V; Carroll & Sapon, 1959) and the Continuous Visual Memory Test (CVMT; Trahan & Larrabee, 1983). From a reminding account perspective of the underlying mechanisms of the distributed practice effect (Benjamin & Tullis, 2010), individual differences in declarative memory should influence the optimal lag for any given to-be-learned item. According to the reminding account, an item gets strengthened more the closer an item is to being forgotten completely (Benjamin & Tullis, 2010). Those with better declarative memory, therefore, will require a wider lag for the to-be-learned item to be closest to total failure, or at a desirable level of difficulty. Those with weaker declarative memory, on the other hand, will require a narrower lag for the to-be-learned item to be at a desirable level of

difficulty. Alternative theories of the underlying mechanisms such as encoding variability and long-term potentiation consolidation processes do not make such claims for a role for declarative memory.

Perhaps surprisingly, few distributed practice studies have investigated the interaction between declarative memory and lag. In one study, Li (2017) conducted an experiment in which eighty-six L1 speakers of English with no experience of learning a tonal language were split into four groups, with ISIs of 1 day or 1 week, and RIs of 1 week (ISI/RI ratio of 14% and 100%) or 4 weeks (ISI/RI ratios of 3.6% and 25%). The participants were taught L2 vocabulary and the oral production of Mandarin tones across five sessions, and they were also given tests of several individual difference measures, including declarative memory as assessed by the Continuous Visual Memory Task (Trahan & Larrabee, 1983). Results showed that while declarative memory played a role in the oral production of the Mandarin tones, there was no interaction between declarative memory and either ISI or RI. To the best of my knowledge, however, no studies have investigated the role of declarative memory in L2 grammar learned under distributed practice schedules.

5.3 The Current Study

5.3.1 Research Questions for Study 2

Given the mixed results regarding previous studies into L2 grammar utilising ISI/RI ratios of 10-30% (Cepeda et al., 2008; Rohrer & Pashler, 2007), and the lack of explicit research into factors that may influence the optimal spacing of L2 grammar, study 2 investigated whether one factor (intentional vs. incidental learning conditions) affects the optimal ISI for a set RI (35 days) when learning the form-meaning connections of an aspect of L2 grammar. In exploratory research, I also included two other factors (items that had previously appeared in the exposure

phase vs. new items that required a generalisation of the rules; and declarative memory) influenced the optimal lag. In this case, the to-be-learned items were artificial determiners that convey both distance from the subject (near or far) and animacy (living, non-living) of the noun. The web-based experiment contained five groups, differing only in the length of ISIs. The research questions for this study were pre-registered, and were as follows:

1. Does distributing the presentation of form-meaning connections (animacy and distance) of determiners produce better results than massed presentation?
2. What is the optimal lag for both intentional and incidental learning conditions of form-meaning mappings (animacy and distance) for a 35-day RI on tests of determiners in a multiple-choice test? Do intentional learning conditions produce a wider optimal lag than incidental learning conditions?

5.3.2 Predictions from Pilot Study

I made a set of predictions (see Table 16 and Figure 12) based on a small set of pilot data estimating the probability of getting a correct binary choice answer in the test phase. The pilot data included just three of the ISI conditions (0-day ISI, n= 10, 2-day ISI, n= 5 , 7-day ISI, n= 8) with the rest of the predictions extrapolated to presuppose a non-monotonic positively- skewed bell curve and a wider optimal spacing for intentional over incidental conditions and memory over generalised items (Janiszewski et al., 2003). I predicted that, in line with previous research (Delaney et al., 2010; Miles, 2014), distributed practice, that is, the optimal ISI for a given condition, would produce better results than massing. Based on the findings from Janiszewski et al. (2003) and Verkoeijen et al. (2005), I predicted that incidental learning conditions would produce a) a narrower optimal ISI than items that have been presented under intentional learning

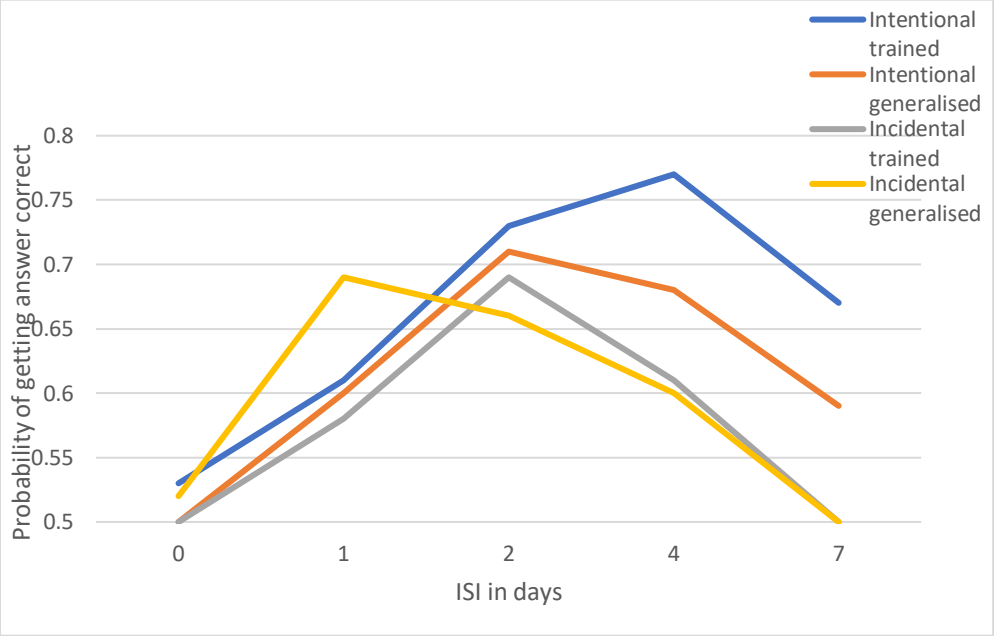
conditions and b) a smaller distributed practice effect, i.e., a lower proportion of correct answers at their respective optimal ISIs.

Table 16. Predictions Based on Pilot Data

Learning condition		ISI in days				
		0	1	2	4	7
Intentional–memory	Probability	.53	.61	.73	.77	.67
	log odds	.12	.33	.87	1.1	.59
Intentional–generalised	Probability	.5	.6	.71	.68	.59
	Log odds	-.12	.078	.021	-.33	-.22
Incidental–memory	Probability	.5	.58	.69	.61	.5
	Log odds	0	.32	.80	.45	0
Incidental-generalised	Probability	.52	.69	.66	.6	.5
	Log odds	.080	.40	-.22	-.12	-.080

n.b. The pilot study collected data from three groups: 0-day ISI, 2-day ISI and 7-day ISI. 1-day ISI and 4-day ISI data is predicted.

Figure 12. Study 2 Predictions based on Pilot Data



Note. This was a small set of pilot data using just three of the ISI conditions (0-day ISI, n= 10; 2-day ISI, n= 5; 7-day ISI, n= 8) with the rest of the predictions extrapolated to presuppose a non-monotonic positively- skewed bell curve and a wider optimal spacing for intentional over incidental conditions and memory over generalised items.

These pilot-based predictions differed in two significant ways from Cepeda et al. (2008), which showed that in an experiment investigating the learning of declarative facts, for a 35-day RI, the optimal spacing was 7 days, and there was a 10% increased chance of getting the answer right compared to the 0-day massed condition. The first difference is that the size of the score gap between massed and the optimal gap differed considerably. In Cepeda et al. (2008) there was a 10% difference between day 0 and day 7 for intentional – memory, whereas for the pilot- based predictions, the difference in scores between day 0 and day 2 for intentional-memory was 20%. The second difference between the Cepeda et al. (2008) predictions and the pilot study data are that the optimal spacing is much narrower for the pilot data (2 days) than Cepeda et al. (2008) (7 days). There are several reasons why I considered that the pilot data may be a more accurate predictor of the optimal ratios of ISI to RI under these conditions than Cepeda et al. (2008). In terms of the strength of the distributed practice effect, the to-be-learned items used by Cepeda et al. (2008) were general knowledge facts, such as knowing the European country which eats the most Mexican food. With such material, it is likely that all participants will remember some facts. Distributing out their study may make it easier, but even those in the massed group will remember some. In order to work out form-meaning connections of determiners, on the other hand, we hypothesised that distributing the presentation makes it easier to abstract the rules in addition to then remembering and retrieving them). If that is the case, then those that do work out the rules would most likely get a much higher score for the incidental condition than those who do not work out the rules, for whom it would remain complete guesswork. Kornell and Bjork (2008), in a study into whether people can abstract the common features of artists' paintings and then recognise their work when presented with previously unseen pieces, found that there was

over a 20% increase in chances of correctly identifying the artist on a one-in-twelve multiple choice question for the spaced presentation condition compared to a massed condition. And in the study 2, even for the intentional condition, in which participants are told the rules before each exposure block, remembering which determiner refers to which meaning is likely to be less memorable after 35 days than an interesting general knowledge fact. Hence, this condition too may have a larger distributed practice effect than in the study of Cepeda et al. (2008). Regarding the optimal spacing differences between the pilot data and Cepeda et al. (2008), the pilot data is closer to the results of Suzuki and DeKeyser's (2017), where their 1-day ISI group outperformed the 7-day ISI group on measures of reaction time. They posited that short ISIs are better for more complex and more procedural tasks and items.

5.3.3 Additional Exploratory Research Questions

In addition to the two main research questions, for which there was adequate power in the study design (see power analysis section below), I included two aspects of data collection that allowed for more exploratory research, that were also included in the pre-registration. Unfortunately, these research questions did not have sufficient power to include as main research questions in the study design but we felt were important to include in order to provide additional data on the effects of generalisation and declarative memory.

3. Is there a distributed practice effect for generalizable as well as trained (memory) items? If so, is there a difference in optimal lag and size of the distributed practice effect?

Based on previous studies (Rohrer & Taylor, 2006; Vlach et al., 2012; Vlach et al., 2008), I predicted that new items that follow the given form-meaning mapping rule but have not been previously presented during the exposure phase (generalised items) would produce a narrower

optimal ISI and a lower overall proportion of correct answers at their optimal ISIs than previously trained (memory) items. See Table 16 and Figure 12.

4. Do individual differences in declarative memory influence the optimal lag?

Based on the reminding account (Benjamin & Tullis, 2010), in which memory traces are strengthened more the closer they are to complete failure upon a second presentation, I predicted that the optimal lag would differ for those with stronger declarative memory compared to those with weaker declarative memory. That is, those with stronger declarative memory would require a wider optimal lag, as this would create a desirable level of difficulty, than those with weaker declarative memory, for whom a narrower optimal lag would create a more desirable level of difficulty.

5.4 Method

5.4.1 Participants

Participants for study 2 were selected via Prolific's (www.prolific.co) demographic filters to pre-screen participants to ensure everyone is an L1 speaker of English, between the ages of 18-60 and university educated. Potential participants who had previously taken part in studies using a similar artificial language were excluded.

A power analysis was conducted and it was determined that 200 participants were needed (mean age = 38.7), of which 114 were female. See Appendix K for the power analysis, which provides a justification for the study design. Recruitment was carried out in accordance with the guidelines of Lancaster University (see Appendix I).

5.4.2 Materials

The artificial language system used in this study comes from Williams (2005) (see also Faretta-Stutenberg & Morgan-Short, 2011; Hama & Leow, 2010; Rebuschat et al., 2013; see Appendix J). Four artificial pre-nominal determiners (gi, ro, ul, ne) indicate both animacy of the object (living, non-living) and distance from the subject of the sentence (near, far). See Table 17. For example, in the sentence

I tried to play dead while *gi* bear sniffed me.

the determiner "gi" denotes both that the bear is living and that it is near. The allocation of determiner to meaning mapping matched Williams (2005) and Rebuschat et al. (2013).

Table 17. The Artificial Determiner System used in Study 2

	Living	Non-living
Near	gi	ro
Far	ul	ne

Forty-eight noun phrases were used in the exposure phase and can be found in Table 18 (for the full list of sentences used in the exposure and testing phases see Appendix L). Of these 48 sentences, 24 sentences were taken from Rebuschat et al. (2013) and I added 24 more to create 12 each of living-near, living-far, non-living-near, non-living-far. New sentences were created following the same guidelines outlined by Williams (2005).

Table 18. The 48 Determiner-Noun Combinations Used in the Exposure Phase and the Test Phase.

	Animate		Inanimate	
Source	Near	Far	Near	Far
Rebuschat et al. (2013)	gi bear	ul bee	ro box	ne book
Rebuschat et al. (2013)	gi dog	ul cat	ro picture	ne plate
Rebuschat et al. (2013)	gi pig	ul fly	ro sofa	ne cushion
Rebuschat et al. (2013)	gi rat	ul bird	ro cup	ne clock
Rebuschat et al. (2013)	gi lion	ul monkey	ro table	ne stool
Rebuschat et al. (2013)	gi cow	ul snake	ro television	ne vase
New	gi lizard	ul mouse	ro bottle	ne umbrella
New	gi eagle	ul penguin	ro coin	ne key
New	gi donkey	ul rhino	ro postcard	ne door
New	gi gorilla	ul chimpanzee	ro toothbrush	ne mirror
New	gi crocodile	ul cheetah	ro battery	ne newspaper
New	gi hedgehog	ul racoon	ro comb	ne bowl

Note. Half the noun phrases are taken from Rebuschat et al. (2013) and half have been created for this study.

For the test phase, there were 96 sentences (see Table 19). Forty-eight sentences were the trained items which tested memory. These were the noun-phrases (determiner plus noun) from the exposure phase, but with a different context. I also created 48 more sentences with new noun-phrases which did not appear in the exposure phase but followed the form-meaning mappings. They were used to test participants' ability to apply the form-meaning mapping rule. These sentences were the equivalent of the new or "true generalisation" items in Rebuschat et al. (2013). The reason for the increase in number of both the exposure and test items over previous versions of this study was to increase the power for testing an increased number of variables that were being investigated (intentional and incidental learning conditions; trained and new items; and declarative memory measures).

Table 19. The 48 Additional, New Determiner-Noun Phrases used in the Test Phase

	Animate		Inanimate	
Source	Near	Far	Near	Far
Rebuschat et al. (2013)	gi elephant	ul camel	ro desk	ne candle
Rebuschat et al. (2013)	gi hamster	ul horse	ro phone	ne lamp
Rebuschat et al. (2013)	gi rabbit	ul turtle	ro spoon	ne towel
New	gi otter	ul squirrel	ro rug	ne bench
New	gi swan	ul beaver	ro bin	ne shoe
New	gi ferret	ul lobster	ro brick	ne knife
New	gi snail	ul deer	ro flag	ne ball
New	gi bat	ul octopus	ro bathtub	ne shovel
New	gi shark	ul dolphin	ro wok	ne paintbrush
New	gi llama	ul kangaroo	ro basket	ne ring
New	gi spider	ul badger	ro spatula	ne corkscrew
New	gi wasp	ul owl	ro pillow	ne bag

5.4.3 Procedure

The experiment was conducted online using Qualtrics (Qualtrics, Provo, UT, USA).

5.4.3.1 Exposure Phase.

Participants were exposed to the 48 sentences once per block for three exposure sessions, each session spaced out with differing lags according to which of five groups the participant has been randomly assigned (0-day, 1-day, 2-day, 4-day, 7-day). The exposure blocks were followed by a delayed test block after 35 days. The ratio of ISI to RI (20%) for the 7-day group was based on Cepeda et al. (2008), who found that a 7-day ISI was optimal for a 35-day RI for a multiple-choice task. As we hypothesised that items that require abstraction of rules, incidental learning conditions and those with poor declarative memory would reduce the optimal spacing, we included four other ISIs, giving five groups in total: 0-day ISI (massed), 1-day ISI with an ISI/RI ratio of 2.8%, 2-day ISI with an ISI/RI ratio of 5.7%, 4-day ISI with an ISI/RI ratio of 11% and 7-day ISI with an ISI/RI ratio of 20%. See Figure 13.

Before each exposure block, participants were given the rules for one aspect of the determiners' form-meaning mappings (intentional) but not for the other aspect (incidental), for

example, that gi and ro indicated “near” and ul and ne indicated “far” but not that gi and ul indicated “living” and ro and ne indicated “non-living”. Participants were then tested on the meanings of the determiners via a two-way forced choice test of the intentional pair of form-meaning mappings. If they answered all four correctly, they moved on to the next stage. If one or more was answered incorrectly, the intentional presentation was repeated and participants were tested again. Once they had all correct, they moved on to the exposure phase. The intentional and incidental conditions were counter-balanced so that for half the participants animacy was intentionally inputted and for the other half distance was intentionally inputted. See Appendix M for the exposure instructions and a sample block.

During the exposure block, participants saw a sentence and pressed a key to indicate whether the object was near or far (if distance was the intentional condition) or whether the object was living or non-living (if animacy was the intentional condition). Participants were told to read the whole sentence, create a mental image of the scene and then make their decision.

5.4.3.2 Test Phase.

In the testing block, the 96 test trials were presented with a two-way forced-choice answer, following Williams (2005) and Rebuschat et al. (2013). For example:
The babysitter poured juice into __ cup for the child.

ro *ne*

Half of the sentences (48) included the same determiner–noun combination encountered in the exposure blocks but with different sentence contexts, and half (48) were new nouns but following the same determiner form-meaning rules. Half the old, trained items (24) and half of the new, rule-governed items (24) tested the intentionally inputted determiners and half (24, old,

trained items and 24 new, rule-governed items) tested the incidentally inputted determiners. See Appendix N for the test block.

5.4.3.3 Debriefing Questionnaire.

After the testing phase, a debriefing questionnaire asked participants questions about which rules they remembered and when they remembered them (see Appendix O). The questionnaire was structured so that the questions became increasing more explicit and more guided in order to avoid, if possible, interfering with their thoughts and interpretations and to provide a clearer understanding of whether the participants were fully, partially or totally unaware of the of the form-meaning connections (see Rebuschat et al., 2015 for a discussion of retrospective reports). Participants were first asked in an open question if they noticed any rules, including any that they had been told at the beginning of the experiment, for the use of gi, ro, ul, and ne. They were also encouraged to write down guesses. For each rule that they wrote about, participants were then asked when they noticed the rule (when the instructions were given, during day 1 practice, between day 1 and day 2 practice, during day 2 practice, between day 2 and day 3 practice, during day 3 practice, between day 3 and the delayed test, during the delayed test, when asked about the rules just now), and their level of confidence (not sure at all, not sure, I think so, I'm very sure). Next, in a more guided question, participants were asked to guess at a rule for the incidental aspect based on the two pairs of determiners. For example, when the intentional aspect was animacy, participants were asked to guess what the rule shared by gi and ro and what the rule shared by ul and ne was. Finally, in the forced guided question, participants were told the incidental aspect of the rule and asked for the incidental meaning of one pair of determiners. For example, when the intentional aspect was animacy, participants were told that

one pair of words referred to near things and the other to far things. They were then asked to choose whether gi and ro referred to near or far things.

The debriefing questionnaire questions relating to what rules were remembered were coded by two researchers, following the coding laid out in Rebuschat et al. (2015). Participants were coded as fully aware, partially aware or unaware. Fully aware participants were those who, in the first question, accurately described in the incidental aspect of the form-meaning connections for each of the four determiners. Partially aware participants were those who mentioned some, but not fully, of the incidental aspect of the form-meaning connection, in either the first open question or the second guided question. Inter-rater reliability was measured using Cohen's Kappa.

Figure 13. Study 2 Study Design

ISI Group													
0-day							P1	P2	P3	35 days RI	Delayed test	Debriefing questionnaire
1-day					P1	1 day ISI	P2	1 day ISI	P3	35 days RI	Delayed test	Debriefing questionnaire
2-day			P1	2 days ISI		P2	2 days ISI		P3	35 days RI	Delayed test	Debriefing questionnaire
4-day		P1	4 days ISI		P2	4 days ISI			P3	35 days RI	Delayed test	Debriefing questionnaire
7-day	P1	7 days ISI		P2	7 days ISI				P3	35 days RI	Delayed test	Debriefing questionnaire

Note. P1, P2, P3 = three presentations of the same 48 sentences. All five groups had an RI of 35 days. The five groups with ISIs and ISI/RI of 0-day; 1-day ISI, 2.8% ISI/RI; 2-day ISI, 5.7% ISI/RI; 4-day ISI, 11% ISI/RI; 7-day ISI, 20% ISI/RI. In the delayed test, 96 test sentences were presented (48 sentences using the noun phrases from the exposure phase, testing memory, and 48 new noun phrases, testing generalisation of the form-meaning mapping rules)

4.4.4 Analysis

Scoring was calculated with one point for the right answer and zero points for an incorrect answer. All data from a participant was excluded before data analyses if the participant: failed both attention checks within one exposure block; or responded to the same side (e.g., pressing the left side button) for 90% or more of responses within a block; or showed a particular alternating pattern (e.g., left/right/left/right) for 90% or more within a block. Due to a technical error, one question in the group for which animacy was the intentional aspect did not register valid answers and therefore was removed from the analysis.

A multivariate logistic regression model (Bates et al., 2015) with fixed and random factors was performed using R (R Core Team, 2019). Fixed effects included the delay group (0, 1, 2, 4, 7 days), item type (generalised and trained/memory) and learning type (intentional and

incidental). The random factors were at participant and item level. The random effects structure was chosen according to the maximum for which the model converged with intercept and slopes. Using the car::Anova() package, the following model was built:

glmer(accuracy ~ (1 + itemtype + learningtype | participant) + (1 + delaygroup | stimuli)
+ delaygroup + learningtype + itemtype + delaygroup:learningtype + delaygroup:itemtype)

For research question 1, if a main effect was found for delay group ($p<.05$), we investigated four contrasts (with an adjusted alpha of $p<.05/4$) to determine whether the delay group has an effect under each of the four combinations of item type and learningtype.

Research questions 2 and 3 can be answered in two ways. Firstly, they can be answered in terms of location, that is, where the peak lies. Secondly, they can be answered in terms of magnitude, that is, whether there is a difference in the likelihood of answering an item correctly in the optimal ISI group compared to the other ISI groups. The former may be more theoretically interesting and meaningful than the latter, and this was addressed using a descriptive analysis of the model-fitted estimates of the likelihood of response accuracy under the given learning condition (intentional vs. incidental) and item type (generalised vs. trained/memory). The difference in magnitude may also be theoretically meaningful to consider, and this was tested using the single model and ANOVA of the main effects of learning condition and item type mentioned above.

Research question 3, which investigated items that were trained compared to items that required a generalisation of the rules, does not meet the conventional threshold for appropriate control of a type 2 error. This means that if a true difference in overall magnitude between

memory/trained items and generalised items exists, this study may not be able to detect the difference. While limitations on available resources are a valid constraint on sample size (Lakens, 2022), this does mean that if no difference in magnitude were detected, further research with a more sensitive study may be warranted. However, type 1 errors were controlled to ensure that the chance of erroneously concluding that an effect exists when one does not is sufficiently limited. Furthermore, while observing a difference in magnitude would be of theoretical interest, as mentioned, of primary concern is the location, that is where the peak of the upside-down u-shape falls. In other words, we are interested in whether the optimal lag is narrower for incidental than intentional learning conditions, and for generalised items than memory-trained items. And this study is adequately designed to assess the location.

In addition to reporting any main effects for learning type (RQ2) and item type (RQ3), we described the optimal delay group for each of the conditions. For research question 2, if there is a main effect for learning type, we then carried out two further contrasts (with an adjusted alpha of $p<.05/2$) to give more detail on whether this main effect was for items that are generalised and again for items that are testing memory. Similarly, for research question 3, if there was a main effect found for item type, we then carried out two further contrasts to determine whether this effect was true for both intentional and incidental conditions. The coding syntax for all of the contrasts was as follows:

```
modelcontrasts <- emmeans(model, ~ delaygroup * learningtype * itemtype)
pairs(modelcontrasts, simple = list("delaygroup", "learningtype", "itemtype"))
```

As each hypothesis test supports a single research question, alpha was fixed at 0.05 as the

threshold for significance. Any post-hoc contrasts were adjusted for family-wise error rates by research question. Power was calculated as the probability of observing the significant effects of interest for each research question, as defined above. See Appendix K. As described in the power analysis section, we did not expect a linear fit through the data, but rather an inverted u-shape, whereby performance increases up to the optimal ISI and then declines afterwards.

Finally, for research question 4, which investigated the role of declarative memory, we added declarative memory to the model. In order to make the model converge, item type was first removed from the model.

The design and analysis of the study was pre-registered, and the pre-registration, all data, and the data analysis scripts can be found on the osf site: osf.io/ j7g54/.

5.5 Results

5.5.1 Massed vs. Delayed

Table 20 details the descriptive statistics for the delayed test scores for intentional and incidental aspects of the form-meaning connections, as well as trained and generalised items at each of the five lag groups. Figures 14 and 15 illustrate the results.

Table 20. Descriptive Statistics for Intentional, Incidental, Generalised and Trained

	0-day		1-day		2-day		4-day		7-day	
	M	SD	M	SD	M	SD	M	SD	M	SD
Intentional	0.6	0.19	0.74	0.21	0.69	0.20	0.75	0.21	0.77	0.19
Incidental	0.52	0.097	0.54	0.082	0.52	0.069	0.52	0.096	0.52	0.054
Trained	0.55	0.11	0.64	0.12	0.6	0.13	0.64	0.13	0.64	0.11
Generalised	0.57	0.11	0.64	0.12	0.61	0.11	0.63	0.12	0.65	0.11

Figure 14. Study 2: Proportion Correct on 35-day Delayed Posttest of Intentional and Incidental Aspect of the Form-Meaning Connections by Delay Group

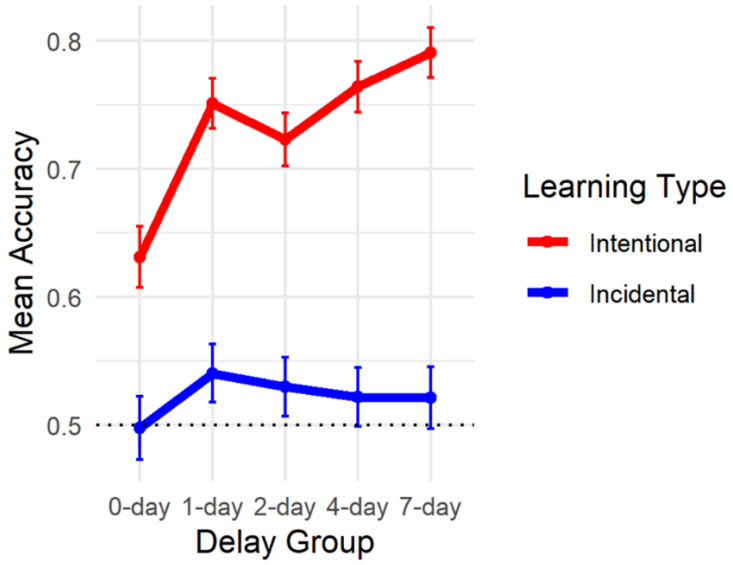
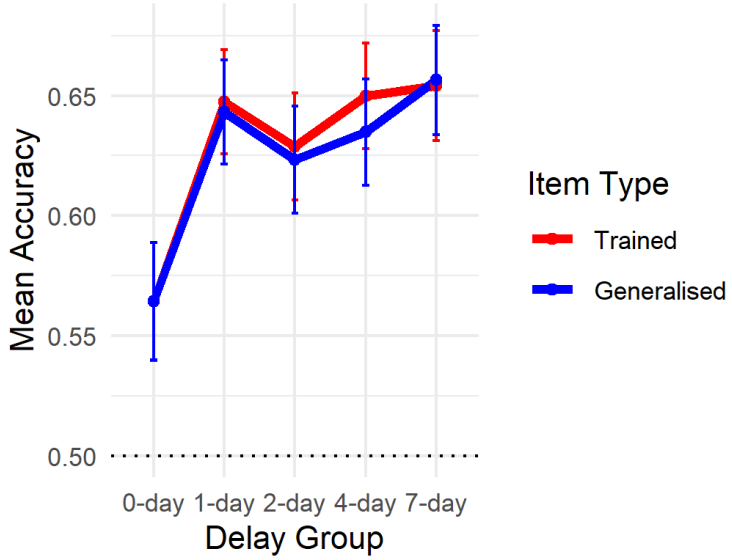


Figure 15. Study 2: Proportion correct on 35-day RI Delayed Posttest when Split by Trained Items that Appeared in the Exposure Phase and New Items that Required a Generalisation of Rules by Delay Group



Analysis of the fixed effects of the model, by means of the Anova () function (type III) in the car package in R (see Table 21), revealed a significant main effect for delaygroup on accuracy, χ^2 (4) = 17.38, p = .002. Contrasts revealed that test scores for the 0-day lag group (M= , SE =, z=,p=) were significantly lower than for the 1-day lag group (M= , SE =, z=-3.277, p=.009), the 4-day lag group (M= , SE =, z= -3.166, p= .013) and the 7-day lag group (M= , SE =, z= -3.340, p= .008) but not for the 2-day lag group (M= , SE =, z= -.2035, p= .25). However, there was no significant difference between the four distributed lag groups (i.e., not the 0-day lag group).

5.5.3 Intentional vs. Incidental

There was also a significant main effect for learning type (intentional vs. incidental) on accuracy, χ^2 (1) = 5.00, p = .025, meaning there was a significant difference between test scores for intentional and incidental aspects of the form-meaning connections. There was also a

significant interaction between delay group and learning type (intentional vs. incidental), χ^2 (4) = 14.86, p = .005.

For the intentional aspect of the form-meaning connection, contrasts revealed there was no significant difference between any of the non-massed groups (1-day, 2-day, 4-day, 7-day lag groups). Regarding the incidental aspect, contrasts revealed that there was no significant difference between the five delay groups for the incidental aspect of the form-meaning connections, suggesting that there was no distributed practice effect. However, test scores for the incidental aspect of the form-meaning connection did reach above chance for the 1-day lag group (M= .54, SD= 0.08, p= .003) and 2-day lag group (M= .52, SD= 0.07, p= .045), suggesting that some learning of the incidental aspect of the form-meaning connection took place in those delay groups.

In order to determine whether the data displayed a similar upside-down u-shape function common to distributed practice data (e.g., Cepeda et al., 2008), in a follow-up analysis, we compared whether a linear or quadratic regression model better fits the relationship between fits the relationship between accuracy and delay group for the intentional aspect of the form-meaning connections. The ANOVA results show that the quadratic model provides a significantly better fit to the data compared to Model 1, the linear model ($F(1, 8684) = 8.1999$, $p = 0.004199$). For the incidental group, on the other hand, a quadratic regression model did not better fit the data than a linear model ($F(1, 8781) = 3.3774$, $p = 0.06613$).

5.5.4 Exploratory Analysis

Regarding the first of the exploratory analyses, investigating whether determiner-noun combinations that had previously appeared in the exposure phase (trained) affected both retention and interactions with distributed practice differently than new determiner-noun

combinations that required a generalisation of the rules (generalised), no main effect was found for item type (trained vs. generalised) on accuracy, χ^2 (1) = .91, p = .34.

Turning to the second exploratory analysis, which investigated whether declarative memory differences between participants affected the optimal lag, there was no main effect for adding declarative memory as a fixed effect to the main model, (*M*=15.91, *SD*=5.30), χ^2 (1) =.29 , p = .591. There was also no interaction between declarative memory and learning type (intentional and incidental), χ^2 (4) = 7.42, *p* = .115. Figure 16 illustrates two scatterplots of declarative memory by delay group for intentional test scores (a), and incidental test scores (b).

Table 21. Summary of the Generalised Linear Mixed-effects Model of (Logs Odds) Accuracy of Response over Delay Group, Intentional vs. Incidental aspects and Trained vs. Generalised:

Note that the intentional, generalised (items that are new and require a generalisation of the rules) with a lag of 0-days (massed) is taken as the reference level for estimation of variation in levels of the factor experimental conditions. R syntax for the final model is as follows: glmer (accuracy ~ (1 + learningtype | participant) + (1 | stimuli) + delaygroup + learningtype + itemtype + delaygroup:learningtype + delaygroup:itemtype), family = binomial)

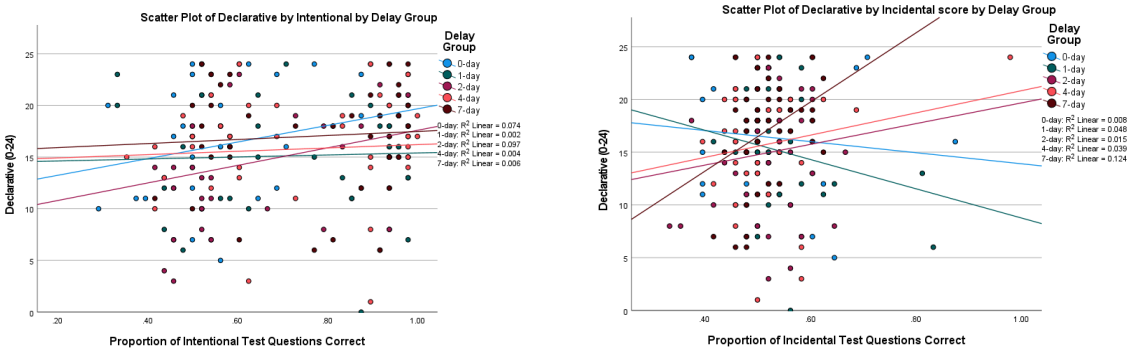
Fixed Effects	Estimated Coefficient	*SE*	Wald Confidence Intervals 2.50%	97.5%	*Z*	*pr(>|z|)*
(intercept)	0.59675	.23658	.084	.9534	2.522	.012
Delaygroup 1-day	1.04550	.33806	.372	1.588	3.093	.002
Delaygroup 2-day	0.67047	0.33231	.026	1.226	2.018	.04
Delaygroup 4-day	1.12957	0.34157	.439	1.667	3.307	<.001
Delaygroup 7-day	1.14341	0.34051	.475	1.701	3.358	<.001
Incidental	-0.55356	0.23675	-.929	-.062	-2.338	.02
Trained	0.04017	0.07187	-.086	0250	0.559	.58
Delaygroup 1-day: incidental	-0.91360	0.33767	-1.454	-.249	-2.706	.007
Delaygroup 2-day: incidental	-0.62653	0.33207	-1.178	.010	-1.887	.059
Delaygroup 4-day: incidental	-1.04367	0.34111	-1.582	-.363	-3.060	.002
Delaygroup 7-day: incidental	-1.11609	0.34008	-1.675	-.456	-3.282	.001
Delaygroup 1-day: trained	--0.06705	0.09997	-.252	.140	-0.671	0.50
Delaygroup 2-day: trained	-0.03155	0.09707	-.215	.165	-0.325	0.75
Delaygroup 4-day: trained	-0.11761	0.10050	-.305	.090	-1.170	0.24
Delaygroup 7-day: trained	-0.04167	0.10086	-.228	.166	-0.413	0.68

Random Effects	Name	Variance	*SD*
Participant	Intercept	2.03857	1.4278
Stimuli	Intercept	0.02428	0.1558

AIC	BIC	logLik	Deviance
22902.6	23051.9	-11432.3	22864.6

Note. 19098 observations, 200 participants, 191 stimuli

Figure 16. Study 2: Scatterplots Showing Declarative Memory by Delay Group



Note. Scatterplots showing declarative memory by delay group for the (left) intentional aspect and (right) incidental aspect of the form-meaning connections.

5.6 Discussion

In this study, I aimed to investigate the optimal lag for form-meaning connections learned under intentional and incidental learning conditions for a retention interval of 35 days. I further investigated whether two other factors, item type (trained items previously studied during the exposure blocks and new items that required a generalisation of the rules) and declarative memory, affected the optimal lag.

5.6.1 Massed vs. Distributed

The first research question was to determine whether distributing the presentation of form-meaning connections (animacy and distance) of artificial determiners produced better results than massed presentation. There was a significant difference in test scores between the 0-day lag group and three of the other lag groups (1-day, 4-day and 7-day, but not 2-day), suggesting that distributing the presentation of the determiners over three sessions on different days was generally more beneficial to learning compared to massing the learning sessions all in one day. This result matched our hypothesis and supports findings by Miles (2014), who found that distributing the practice of L2 grammar resulted in significantly better long-term retention

than massing. It also supports findings in the recent meta-analysis into distributed practice in L2 language learning by Kim and Webb (2022), who found that distributed practice was more beneficial than massed. However, Kim and Webb found that spacing effects were statistically unstable regarding L2 grammar learning. Therefore, the results of the current study are useful in adding to the cumulative evidence in favour of spacing for L2 grammar compared to massing.

Most theories of the underlying mechanisms of the distributed practice effect can explain the finding that distributed exposure was significantly better than massed exposure of the determiners. From a deficient processing perspective, less attention paid either automatically or consciously to the to-be-learned items resulted in less encoding. An encoding variability explanation suggests that the more varied contextual cues encoded alongside the to-be-learned items on different days compared to those encoded in the massed group, which were presumably the same throughout the three phases of the study, were more likely to overlap with the contextual cues encoded during the testing phase. From a reminding account perspective, the desirable level of difficulty of retrieval may have been too easy for the massed group. And from a consolidation account, the massed group would not have benefitted from the time for the to-be-learned items to stabilise and strengthen between exposure blocks, including periods of sleep. While the results of research question 1 do not allow us to discriminate between the different accounts of the underlying mechanisms, the results for research questions 2 and 3 discussed below offer more insight into the underlying mechanisms,

Methodologically, the finding that the distributed groups performed significantly better on the delayed posttest than the massed group, at least for the intentional aspect of the form-meaning connection, is instructive in terms of how distributed practice studies are designed and interpreted, particularly around how the term *massed* is used for both within one day and ISIs

over 1 day or more. While Suzuki and DeKeyser (2017a) used the term *massed*, when they found that there was no statistical difference between massed and distributed practice of L2 morphology, they were in fact comparing longer versus shorter lags. Their so-called massed group had an ISI of 1 day. It would be interesting to replicate their study with an addition massed group, with an ISI of 0 days. Another example of different interpretations of the same labels is the massed group in the contextual vocabulary learning study by Elgort and Warren (2014), which involved reading a chapter with the target vocabulary repeated within. The authors reported that a number of participants read a chapter over two days in the massed group. This leaves open the question as to what extent the distribution of reading practice and incidental learning vocabulary over 24 hours affected their results. These two examples highlight the importance of keeping the term massed for 0-day ISIs and if it is within the same day, clearly reporting the ISI.

5.6.2 Intentional vs. Incidental

The second research question sought to determine the optimal spacing for both intentional and incidental form-meaning mappings for the 35-day retention interval, and to thus investigate whether intentional learning conditions require a wider optimal lag than incidental learning conditions. For the intentional aspect of the form-meaning connections the delay-group with the highest accuracy scores on the delayed posttest was the 7-day group (M = .79), which is a 25% improvement on the 0-day lag, massed group. However, there was no significant difference in test scores for the intentional aspect of the form-meaning connections among the four non-massed test groups (1-day, 2-day, 4,-day and 7-day lag), suggesting that there was no one optimal lag. While this result runs contrary to my predictions, it aligns with Kasprowicz et al. (2019), who found no advantage on a 42-day RI for either a 7-day lag (16.7% ISI/RI ratio) or

a 3.5-day lag (8.3% ISI/RI ratio) for children learning French verb inflection. Cepeda et al. (2008), in their study investigating the learning of interesting facts, found that for a 35-day RI, the optimal lag for recognition tests was a 7-day lag (20% ISI/RI ratio), and that their data fit an upside-down u-shape function of performance (see Maddox, 2016 for other studies that demonstrate this function). However, crucially, they made no claims that their 1-day or 4-day lag groups achieved significantly worse test results than the 7-day lag group. Instead, through fitting a cubic spline, they showed that the 7-day lag was the peak of the upside-down u-shape, that is, the optimal lag. Indeed, post-hoc analysis on their data (Wiseheart, personal communication, September 2023) shows that the scores on the delay test for the 1-day, 4-day and 7-day lag groups were not significantly different from each other. The finding in the current study that a quadratic regression model fits the intentional data better than a linear model suggests that the 7-day lag could possibly be the peak of the upside-down u-shape curve. However, without lags at greater lengths (including, for example, a 14-day lag group), it remains speculative.

Even if there is an optimal lag which is at the peak of the upside-down u-shape function, the results of the current study suggest that for the intentional aspect, lags between 1-day and 7-days produce similar results on a 35-day delayed test, and that it does not matter so much where that optimal lag may lie. Wiseheart (personal communication, September 2023) suggests that while there may technically be an optimal peak, for practical, including pedagogic, purposes, it may be more constructive to think about it as an optimal *window*, within which performance is comparable. However, from a theoretical point of view, it is useful to consider why performance on the delayed posttest is equally effective when exposure is spaced at any point between one day and seven days.

One explanation for the findings that there was no difference between the four spaced groups is that there may be more than one underlying mechanism at work. The reminding account (Benjamin & Tullis, 2010) predicts that there is a lag at which there is an optimally desirable level of difficulty involved in retrieving the previously presented items. At sub-optimal lags, the difficulty level is either too high or too low, resulting in sub-optimal performance. However, Kornmeier et al. (2012) postulated that there may be more than one optimal lag, or at least peak, for any one retention interval. To support these hypotheses, Kornmeier et al. cited the meta-analysis of spacing studies by Cepeda et al. (2006), who found that a 1-day lag was the most beneficial lag for a number of different retention intervals ranging from 1-day (ISI/RI ratio of 100%) to 28 days (ISI/RI ratio of 3.6%). Kornmeier et al. (2012) hypothesised that this may be related to long-term potentiation at a synaptic level. That is, when a synapse is stimulated, an increase in synaptic efficacy occurs. Three phases of long-term potentiation have been differentiated, depending on the stimulation frequency, with decay times in hours, days and months respectively (Abraham, 2003). In a follow-up empirical study, Kornmeier and colleagues (Kornmeier et al., 2014) used a paired-word learning task (German-Japanese word pairs) and divided participants into six groups according to lags ranging from 7 minutes up to 24 hours. Participants were presented with and then tested on the word pairs in five separate blocks and then tested in three posttests (24-hours, 7 days and 28 days). Results suggested one peak around 12 hours and another at 20 minutes. While this study did not include groups with longer lags and is therefore difficult to compare with the current study, it does provide some supporting evidence for the presence of more than one optimal lag for any given retention interval.

In addition to the reminding and desirable difficulty mechanisms which could kick in after a number of days, another possible mechanism, and one which may result in comparable

spacing advantages at 1-day lags, is sleep consolidation (Bell et al., 2014), which predicts an enhanced benefit for periods that include sleep. In a paired associates experiment by Bell et al. (2014), results on the 10-day RI test revealed similar scores for the 12-hour lag with sleep group compared to the 24-hour lag with sleep group but an advantage for the 12-hour lag with sleep group over the 12-hour lag without sleep group. In the study by Kornmeier et al. (2014), the authors admitted that the 12-hour lag group benefited from at least one episode of sleep during the experiment, and that this may have resulted in peak performance, due to the combination of sleep consolidation and the comparatively less forgetting in the 12-hour group than in the 24-hour lag group. Turning back to the results from study 2, it is possible that an initial large benefit from sleep consolidation after 24 hours, and/or the effects of different phases of long-term potentiation exerting influence at different times, provide this early peak in performance at 1-day lag, with a further peak occurring when there is a desirable level of difficulty for memories to be successfully retrieved. And this, in turn, may explain the lack of significant difference between the different lag groups.

Moreover, this potential role of consolidation resulting in an early peak around the 1-day lag may be amplified by the type of item that is to be learned. Items that are thought to rely more on procedural memory, such as L2 grammar, may, as Kim et al. (2013) postulated, interact differently with distribution of practice compared to items that rely on declarative memory. Kim et al. (2013) hypothesised that only aspects of the to-be-learned item that rely on declarative memory benefit from distributed practice, as procedural memory has much lower forgetting rates. Procedural memory, on the other hand, may benefit from offline consolidation resulting in gains during training, and perhaps more so over the first 24 hours (Simor et al., 2019). L2

grammar has been linked with declarative memory at early stages of acquisition (Hamrick, 2015) and also with procedural memory (Ullman, 2001).

With regard to the learning condition, there was a significant difference in intentional and incidental test scores and a significant interaction between learning condition and lag. However, this does not really support the findings by Janiszewski et al. (2003), who found that intentional learning conditions resulted in a stronger distributed practice effect and a wider optimal lag than incidental learning conditions. The reason for this is that there was little evidence of a distributed practice effect for the incidental aspect of the form-meaning connections. While the regression model shows that the 1-day and 2-day groups were significantly above chance, there was crucially no significant difference between any of the delay groups. That is, massing the presentation of the incidental aspect of the form-meaning connection was just as effective as distributing it.

There are several possible explanations for why learning of the incidental aspect of the form-meaning connections did not show a distributed practice effect in terms of significant differences in test scores between the different delay groups.

One possible explanation relates to initial encoding. It may be possible that, rather than a dampened down distributed practice effect with all scores close to chance due to a failure to maintain the incidentally inputted form-meaning connections over the 35 days to the delayed posttests, the to-be-learned material and learning activities were too complex to be initially encoded under incidental learning conditions. In Kim and Webb's (2022) meta-analysis of distributed practice in L2 language learning, they found a number of moderators that appeared to reduce the effect of spacing. These included: grammar instead of vocabulary, comprehension activities instead of paired-associate, study only trials instead of test trials. Study 2 included

each of these moderators. As a result, it is possible that the incidental aspect of the form-meaning connections was not encoded in the exposure block to a sufficient extent to allow it to subsequently be maintained over 35 days. However, results from Rebuschat et al. (2013) and Williams (2005) show that encoding of the incidental aspect of the form-meaning connection (animacy) is possible. By way of comparison, the study by Rebuschat et al. (2013) used the same artificial determiner system as the current study and gave the exposure blocks one after the other as in the massed (0-day lag) group in the current study. However, they gave an immediate posttest rather than the delayed posttest with an RI of 35 days as used in the current study. Rebuschat et al. found that participants scored 74.8% of the 2WFC (incidental aspect) questions correctly. Similarly, 69% of participants in Rebuschat et al. were able to verbalise some knowledge regarding the incidental aspect of the form-meaning connection. This suggests that initial encoding is unlikely to be the primary reason for the lack of distributed practice effect.

If the incidental aspect of the determiners and the context in which it was presented was too complex to learn, the above chance scores for the 1-day and 2-day lags could possibly be explained by awareness. That is, becoming aware of the incidental rules, either fully or partially, may have resulted in the marginally above chance learning in 1-day and 2-day lag groups. With so few participants becoming aware of the incidental aspect of the form-meaning connections (5 could fully articulate the rules, a further 10 demonstrated a partial understanding), it was clearly a difficult task after 35-days. Re-analysis of the data with participants who were either fully or partially aware of the incidental aspect of the form-meaning connection removed shows that 1-day lag ($M = .52$, $SD = 0.05$, $p = .08$) and 2-day lag group ($M = .52$, $SD = 0.07$, $p = .06$), were no longer above chance). This adds credence to the suggestion that over a 35-day RI, it is awareness that drives retention.

A second potential explanation relates to a problem with maintenance of memory traces. It is possible that a 35-day gap to the delayed test was too long for participants to maintain knowledge of the incidental aspect of the form-meaning connections, no matter whether the exposure blocks were distributed or massed. As mentioned above, participants in Rebuschat et al. (2013) scored 74.8% of the incidental questions correctly and 69% of participants were able to verbalise some knowledge of the incidental aspect. Whereas, in the current study, after 35 days, only 52.3% of the incidental aspect 2WFC test questions were scored correctly (51.6% in the massed group). Indeed, in the current study, only 61% of participants could recall some knowledge regarding the intentional aspect of the form-meaning connection, that is, the aspect that participants were explicitly instructed on at the beginning of each of the three exposure blocks. Moreover, only 8% of participants were able to do verbalise some knowledge of the incidental aspect of the form-meaning connection. In order for the knowledge of the form-meaning connections to be maintained for 35 days, it may be that more exposure blocks were needed. Alternatively, in their meta-analysis of explicit vs. implicit instruction, Goo et al. (2015) found that on long-delayed posttests, operationalised as over 30 days, implicit instruction only had a small effect size ($g = 0.345$) compared to a medium to large effect for explicit instruction ($g = 0.747$). It may be possible, therefore, instead of an upside u-shape function of the distributed practice effect peaking between 1-day and 2-day lags as I had predicted, the 35-day RI dampened down the test scores for the incidental aspect in all delay groups towards chance at .5, resulting in no significant difference between the two delay groups that were significantly above chance and the other three delay groups.

It is interesting to speculate the extent to which online recruitment, and the associated study design, affected the results of the incidental learning aspect of the form-meaning

connections. Evidence for online recruitment affecting the results for the incidental aspect comes from the findings that the incidental test results for study 2 were considerably lower than the pilot data results. For example, for the two-day lag group, the mean score for the incidental learning aspect of the form-meaning connections was .52 for the main study 2 results but .69 for the pilot data results. Of course, there were only ten participants in the pilot study and therefore the results may well be skewed. However, while both sets of data were collected via Qualtrics (Qualtrics, Provo, UT), only the main study 2 participants were recruited via Prolific (www.prolific.co). Many Prolific users complete studies as a significant source of income and therefore aim to complete the studies as quickly as possible. It is possible that during the exposure phase, participants could complete the exposure task, particularly when animacy was the intentional aspect, by choosing living or non-living without reading the whole sentence. That is, the participant might only need to read as far as the determiner and noun to decide whether it was living or non-living. Take the following example.:

The researcher studied ul bee from a safe distance.

A participant would only need to read as far as "ul bee" to ascertain that the noun was living and would therefore not take in whether the prepositional phrase was "from a safe distance (far) or, say, with "under a microscope" (near). And therefore,, even though a participant may score 100% on the exposure task, if the participant did not read the whole sentence, meaning attention at the level of detection, it would be almost impossible to incidentally acquire the distance aspect of the form-meaning connection, as the distance notion was embedded within the rest of the sentence (Schmidt, 2010). Unfortunately, reaction time was not measured at an individual trial level, so it is difficult to remove participants who potentially did not read the whole sentence. However, for the group who had distance as the intentional aspect of the form-meaning

connection, it was less easy to complete the exposure task without reading the whole sentence as the distance aspect is embedded more within the context of the sentence. Therefore, one way to negate the possibility of participants not reading the full sentence would be to look at just the half of the participants who had distance as the intentional aspect (see Table 22). While there was no significant difference in scores between those that had distance as the intentional aspect and those who had animacy as the intentional aspect, there was a difference in terms of scores that were significantly above chance. For those who had animacy as the intentional aspect, and who therefore were exposed to distance as the incidental aspect, no groups were significantly above chance, although 1-day and 2-day groups approached it. For those who had distance as the intentional aspect, on the other hand, the 1-day, 4-day and 7-day lags were all significantly above chance. This suggests that perhaps the combined results might be underreporting the incidental learning and retention, particularly of the animacy aspect. However, just including the participants that had distance as the intentional aspect of the form-meaning connections, there was still no significant difference between the different lag groups for the incidental aspect of the form-meaning connection.

Table 22. Proportion of incidental aspect of form-meaning connection test answers correct, organised by which aspect of the form-meaning connection was intentionally inputted.

Intentional	0-day	1-day	2-day	4-day	7-day
Animacy	.53	.52	.52	.50	.50
Distance	.50	.57 **	.53	.55 *	.53 *

*p> .05; ** p > .01

One way to address the challenge of encouraging participants to process the whole sentence in a future study would be for the exposure sentences to be aural instead of written, with participants only allowed to make their choice after they had heard the whole sentence.

5.6.3 Generalised vs. Trained

For the first of the exploratory research questions, I investigated whether new determiner-noun pairings that required a generalisation of the rules benefitted from a distributed practice effect in addition to trained determiner-noun pairings that had appeared in the exposure blocks. For the intentional aspect of the form-meaning connections, generalised test items benefitted from a distributed practice effect. As with the trained test items, there was a significant difference between the 0-day lag group and the 1, 4 and 7-day lag groups (but not the 2-day lag). That is, distributed practice was significantly better than massed practice. This finding adds to previous research that suggests that distributing practice promotes generalisation better than massing (Vlach et al., 2008; Vlach et al., 2012).

Regarding the difference between generalised and trained items, my hypothesis was not met: there was no significant difference between trained and generalised test items. This finding suggests that, for the intentional aspect of the form-meaning connections, perhaps it was not the individual determiner-noun pairings that were remembered after 35 days, but rather the "rule". Indeed, the debriefing questionnaire data revealed just how difficult it was for participants to remember the intentionally inputted rule, with many forgetting completely (39%) or only partially remembering (29.5%) them. It is possible, meanwhile, that the individual determiner-noun pairings were forgotten over such a long period of time.

For the incidental aspect of the form-meaning connections, again there was no difference between the trained and the generalised test items. As the incidental aspect was not explicitly

taught, remembering the rule cannot be the reason. However, given that the there was no statistical difference between trained and generalised test items, it is unlikely that individual determiner-noun pairings were remembered either. Instead, it may be possible that a partial understanding of the form-meaning connection for the incidental aspect of the determiners was acquired at an implicit, or pre-verbalisable level.

5.6.4 Declarative Memory

For the second exploratory research question, I investigated whether individual differences in declarative memory influenced the optimal lag. No significant main effect was found for declarative memory and no interaction was found for declarative memory and delay group. This was somewhat surprising. The first reason for this is that differences in declarative memory, which is a long-term memory system believed to be responsible for episodic and idiosyncratic items (Ullman, 2001), might be expected to affect the retention of the mini-determiner system after 35 days. Put simply, if you have a better memory, you are going to do better when tested after more than a month without study and if you have a worse memory, you are going to do worse. In study 1, declarative memory, measured via MLAT-V, predicted results on several aspects of the artificial language, including word order under massed conditions and adjectives and case markers under distributed conditions.

The lack of significant interaction between declarative memory differences and distributed practice provide the second surprise. From a reminding account (Benjamin & Tullis, 2010) perspective, a desirable level of difficulty, or the point when a to-be-learned item is close to catastrophic memory failure and therefore strengthens the most on retrieval, might require a wider optimal lag for those with stronger declarative memories and a narrower optimal lag for those with weaker declarative memories. The lack of interaction does not support a reminding account explanation of the underlying mechanisms, and instead other theories that do not predict a role for declarative memory in optimal lags may better explain it. Possibilities include an encoding variability account (Glenberg, 1979), and reconsolidation accounts (Smith & Scarf, 2017). However, before ruling out the reminding account, it is worth considering other reasons

for the lack of main effect and interaction between declarative memory and lag. It is possible, for example, that the online data collection adversely affected results.

One aspect of the results hints at interactions between declarative memory, incidental learning and lag. The scatterplots of declarative memory by delay group for intentional and incidental learning (see Figure 16) shows that for the incidental aspect of the form-meaning connections, there appears to be a (non-significant) contrasting interaction. For those in delay groups with relatively longer lags (7-day), having better declarative memory resulted in better scores for the incidental aspect of the form-meaning connection. For those in delay groups with narrower lags (0-day, 1-day), having better declarative memory resulted in worse scores for the incidental aspect. This was only for the incidental aspect. It is not clear why this might be. It is possible that strong declarative memory helps for a desirable level of difficulty at longer lags. Whereas, at shorter lags, stronger declarative memory may shift attention away from incidental aspects of the form-meaning connections or keep them on the intentional aspect. However, a study that specifically investigates potential contrasting aptitude by learning condition interactions would be useful.

5.7 Summary of Study 2

In study 2, I investigated factors that might influence the optimal lag for a given retention interval (35-day RI) for form-meaning connections (animacy and distance) of artificial determiners. I explored whether the learning condition (intentional or incidental), item type (trained determiner-noun combinations that had previously appeared in the exposure phase or new combinations that required a generalisation of the rules) and individual differences in declarative memory. Results indicated that distributed exposure resulted in better results than massed exposure. Results for the intentional aspect of the form-meaning connections were significantly higher than those for the incidental aspect. For the intentional aspect, there was not

one optimal lag. Test results for the four distributed groups (1-day, 2-day, 4-day and 7-day ISI groups) were not significantly different from each other, suggesting that, rather than on peak, there is an optimal window in which, pedagogically, it does not matter whether a lag is one day or seven. The incidental aspect did not show a distributed practice effect. That is, there was no significant difference between any of the delay groups. However, two groups (1-day and 2-day ISI groups) did demonstrate learning at significantly above chance levels. There was no significant difference between test results for trained and generalised items. This suggests that rules were maintained and retrieved over 35 days rather than individual determiner-noun combinations. There was no significant effect for declarative memory on delayed test results. While this was surprising, there was some minimal evidence that there may be complex three-way interactions between declarative memory, with lag and learning condition.

In the following chapter, I synthesise the findings from both studies and discuss their implications for our understanding of the distributed practice effect in L2 learning. This discussion will cover several areas, including the comparison of massed vs. distributed learning, different lags and ISI/RI ratios, different types of knowledge, individual differences, and the insights these findings provide into competing theories explaining the mechanisms behind the distributed practice effect. Finally, I will propose an alternative model to Kim et al.'s (2013) skill retention theory.

Chapter 6: General Discussion

6.1 Introduction to General Discussion

The present thesis reported the results of two studies that investigated the role that the distribution of the practice and presentation of L2 lexis and grammar play when learned under incidental learning conditions.

The aims of study 1 were to find out whether: i) adult learners can acquire a complex artificial language consisting of verbs, nouns, adjectives and case markers bound by a SOV/SVO syntax through a cross-situational learning paradigm; ii) learning was durable after 24 hours; iii) distributing the cross-situational learning exposure increased learning and retention compared to massing it; and iv) five individual memory differences (visual and verbal declarative memory, procedural memory, phonological short-term memory and working memory capacity) predicted the cross-situational learning process, and how they interacted with distribution of learning. Study 2 looked at the learning of form-meaning connections at more educationally relevant time scales with the retention interval set at 35 days. The aims for study 2 were to find out whether: i) distributing exposure to form-meaning connections resulted in better retention after 35 days than massing; and ii) several factors (intentional and incidental learning conditions, generalised vs. training items, and declarative memory) influenced the optimal lag for the acquisition of the form-meaning connections.

In the following sections, I will draw together the findings from the two studies and discuss the implications for our understanding of the distributed practice effect for L2 learning, centred around several areas, including massed vs. distributed L2 learning, different lags and ISI/RI ratios, complexity / declarative and procedural knowledge, the role of individual differences, and the support the findings offer to competing theories of the underlying

mechanisms of the distributed practice effect. I will finally offer a potential alternative model to the skill retention theory of Kim et al. (2013).

6.2 Massed vs. Distributed Practice

One of the primary aims of both studies in this thesis was to investigate whether distributing L2 learning is more beneficial than massing. The findings from the two studies were inconclusive. For study 1, a cross-situational learning study in which participants learned a comparatively complex artificial language with nouns, verbs, adjectives and case markers did not find a benefit for distributed practice over massed. Study 2, which focused on the learning of four determiners that conveyed both animacy and distance, found an advantage for distributed practice over massed for the intentional aspect of the form-meaning connections, but not for the incidental aspect. There are several potential reasons for this difference.

The first and most obvious possible explanation for the differing findings between massed and distributed groups in the two studies involves the learning condition. Study 1 was conducted under incidental learning conditions in that there was no explicit instruction of the rules of the syntax or form-meaning mappings of the vocabulary. In study 2, only the aspect of the form-meaning connections that was intentionally inputted resulted in a distributed practice effect. The incidental aspect, on the other hand, demonstrated no significant difference between the massed and distributed groups. However, previous studies have found distributed practice effects for incidental learning conditions (Challis, 1993; Glenberg & Lehman, 1980; Greene, 1990; Greene & Stillwell, 1995; Jensen and Freund, 1981; Shaughnessy, 1976; Verkoeijen et al., 2005), so, while it may be one factor that influenced the result, it is likely that there are also other contributory factors.

One of those other potential reasons for the different results between the two studies regarding massed and distributed practice might be the retention interval, the length of the lag and the ratio of ISI to RI used in each study. In study 1, the retention interval was 24 hours and the lag for the distributed group was short, only 20mins, giving an ISI/RI ratio of 1.3%. For study 2, the retention interval was a much longer 35-days, with lags ranging from 1-day (ISI/RI ratio of 2.9%), 2 days (5.7%), 4 days (11.4%) and 7 days (20%). However, the shortest lag for study 2 was not far off the ISI/RI ratio of study 1. It is possible, albeit unlikely, that had the distributed group in study 1 had an exactly equivalent ISI/RI ratio to the day-1 ISI group in study 2, that is, an ISI/RI ratio of 2.8% or an approximate 40-minute lag, the distributed group in study 1 may have also achieved significantly higher scores than the massed group on the delayed posttest. In terms of length of the lag, the period of sleep between study sessions in study 2 for the narrowest lag group (1-day ISI, ISI/RI ratio of 2.9%) may have benefitted learning where the absence of sleep in study 1 (20-min lag, ISI/RI ratio of 1.3%) did not. That is, sleep consolidation processes may have been beneficial during training. Lastly, the retention interval, separate from the ISI/RI ratio, may have caused the different results between the two studies. Clearly, memory traces will degrade more over 35 days than 1 day. The lack of strong encoding on the more incidental study 1, may have resulted in less retention had it been over 35 days like study 2. Methodologically, while two lag groups (study 1) or five lag groups (study 2) for one given retention interval helps determine the role that lag on its own and ISI/RI ratio play, studies that include two different retention intervals (e.g., Kasprowicz et al., 2019) allow for retention interval to be considered as a fixed factor on its own, thus allowing for the separation of retention interval from lag.

Another possible explanation for the different findings between the two studies regarding distributed and massed groups involves the type of knowledge learned and the type of task performed. Study 1 involved learning both lexical items and their syntactic roles within the sentence, whereas study 2 involved learning the form-meaning connections of determiners, embedded within an English (and therefore understandable) sentence. Participants in study 2, therefore, were not required to work out the syntactic position of the determiners as they could see the determiner's place after verbs and before nouns. Also, they were told in the instructions at the beginning of the experiment that the determiners had the same meaning as "the". That is to say, the mini artificial language in study 1 was arguably more syntactically complex than the determiners in study 2. It is possible, therefore, that the added syntactic complexity resulted in worse encoding, and thus, there was less benefit for distribution of exposure.

6.3 The Optimal Lag and ISI/RI Ratio and Factors that might Influence them

One of the aims of study 2 was to investigate the optimal lag for a given retention interval and to investigate factors that might influence the lag. In study 2, the results of the intentional aspect of the form-meaning connections suggested that there was not one optimal lag for the 35-day retention interval. Results on the posttest were not significantly different in the 1-day (2.9% ISI/RI ratio), 4-day (11%) and 7-day (20% ISI/RI ratio) lags. Instead, while there may technically be an optimal peak, that is a highest point, as Wiseheart (personal communication, September 2023) suggests, an optimal window may be a more appropriate way to look at spacing. The incidental aspect of the form-meaning connections, while reaching significance, did not demonstrate a significant difference between the massed group and other distributed lag groups, indicating that there was not a distributed practice effect. This was in line with the

findings from study 1, which also involved incidental learning conditions, in not showing a difference between massed and distributed groups.

Incidental learning conditions are unlikely to follow the same ISI/RI ratio as the 10-30% optimal range laid out by Rohrer and Pashler (2007), drawing as they did on the paired-associate (intentional) study results of Cepeda et al. (2008). With less attention given to the to-be-learned items, and weaker representations encoded in memory, incidental learning at early stages under immersion settings may either benefit from massed practice or from much narrower ISIs than the 10-30% optimal ratio range (Rohrer & Pashler, 2007). Indeed, it may be that the optimal ISI/RI ratio is itself an oversimplification. Factors that influence differences in encoding, maintenance and retrieval of to-be-learned items may in turn influence the length of possible retention both during lags and to the delayed posttest. That is, certain items that are not encoded strongly may not be retained over longer retention intervals no matter what the ISI/RI ratio.

Items that required a generalisation of rules did not show a different optimal lag than items that had been presented in the exposure phase. As discussed in chapter 5, this suggests that it is the rule that is maintained over 35 days rather than the individual determiner-noun collocations. It remains to be seen whether a shorter RI, which may produce more of a distributed practice effect under incidental conditions, would result in similar retention for generalised and trained items.

Declarative memory did not influence the optimal lag, which is inconsistent with a reminding account (Benjamin & Tullis, 2010). Individual differences will be discussed in the following section.

6.4 The Role of Individual Differences in Distributed Practice

How individual differences interact with massed and distributed learning schedules is an area of growing interest (Knabe & Vlach, 2020), not least because it can aid our understanding of the mechanisms of the distributed practice effect.

Individual memory differences interacted with distributed practice schedules under incidental learning schedules in study 1, although not in the way that we had predicted. The distributed group was influenced by more individual memory difference measures than the massed group. For the delayed test 5 on day 2, only word order had a predicting variable in the massed group (declarative memory), while in the distributed group, nouns (procedural memory), adjectives and case markers (both declarative memory) had predicting individual difference measures. On day 1, too, results on the lexical and word order tests were predicted by more individual difference measures in the distributed group than the massed group. As discussed in chapter 4, this could be the result of an undesirable level of difficulty forcing a reliance on certain individual difference measures (DeKeyser, 2012).

Declarative memory was an individual difference that did not interact with distributed practice in the same way across the two studies. In study 1, as mentioned above, declarative memory predicted success with adjectives and case markers in the distributed group and word order in the massed group. In study 2, however, no interaction between declarative memory and learning or distributed practice schedules was found. In chapter 5, the possible reasons for the lack of influence of declarative memory on learning and retention, and the lack of an interaction between declarative memory and distributed schedules were discussed, including the potential design problems associated with the online study. Further research, including a partial

replication of study 2, is needed to investigate how declarative memory interacts with distributed practice.

In study 1, individual differences in procedural memory predicted success on the learning of the cross-situational learning task during the training phase in the distributed group only. This supports previous research that has found benefits for distributed practice for procedural tasks during training, albeit in non-verbal tasks (e.g., Lee & Genovese, 1988).

One of the more intriguing findings from study 1 is the differing interaction between individual difference measures and distributed practice, in which those with strong declarative memory and phonological short-term memory did better on verb tests during training if they were in the massed group but worse if they were in the distributed group. Whereas those with strong declarative memory and phonological short-term memory did better on adjective test and case marker tests respectively if they were in the distributed group and worse if they were in the massed group. Distributed exposure may result in a shift in attention for those with strong declarative memory, either due to strong declarative memory increasing the strength or quality of encoding or reducing forgetting through the lag. Alternatively, a shift in attention, caused by unknown other factors, may result in a reliance on declarative memory.

6.5 Competing Theories of the Underlying Mechanisms of the Distributed Practice Effect

The two studies in this thesis did not explicitly test the competing theories of the underlying mechanisms of the distributed practice effect, and as such, unsurprisingly, no clear winner emerged from the data.

Given that there was not one optimal lag for the 35-day retention interval for the intentional aspect of the form-meaning connection, this hints at more than one mechanism

playing a role, and potentially for different mechanisms playing roles under different learning conditions and for different types of tasks. The finding that declarative memory did not interact with lag in study 2 is inconsistent with a reminding account (Benjamin & Tullis, 2010). However, it is interesting to note the extent to which declarative memory predicted success in the distributed group in study 1. Further research is needed on the extent to which declarative memory interacts with lag, and to what extent visual declarative memory overlaps with visualspatial working memory. Other aspects of the findings of study 2, such as the significant improvement in model fit for a quadratic model compared to a linear model for the intentional aspect of the form-meaning connections, are consistent with not only a reminding account (Benjamin & Tullis, 2010), but could also apply to dual-mechanism accounts that include study-phase retrieval and contextual variability (Verkoeijen et al., 2004) and reconsolidation accounts (Smith & Scarf, 2017).

Study 2 results, with the relative peak in scores for the intentional aspect around 24 hours, were consistent with reconsolidation accounts (Smith & Scarf, 2017), with benefits for offline learning over periods of sleep. Sleep consolidation appeared to play a role in the results of study 1. While the lack of significant difference between massed and distributed groups on the test results of the four lexical categories and word order suggest that there was little difference in terms of the amount of consolidation during between massed and distributed groups, there was evidence for different consolidation rates overnight. Overnight consolidation rates differed according to the aspect of the artificial language and whether the learning schedule was massed or distributed. Test results for case markers improved significantly overnight in the massed group but only increased insignificantly in the distributed group, a difference that approached significance. Word order test results between massed and distributed groups also

approached significance, with the distributed group improving overnight but the massed group worsening. The question then is why a 20-minute lag, or lack thereof, affects the extent to which an item is consolidated overnight. One possible explanation from a consolidation account there is a qualitative difference in the memories that are stored after distributed practice, even if there is no quantitative difference as measured by tests during training (Abraham, 2003; Blis & Collingridge, 1993). These results also raise the possibility of different types of linguistic knowledge (e.g., grammar, vocabulary) interacting with distributed practice differentially to result in different amounts of consolidation. And while the lags in study 1 were only 20 minutes, the results suggest that with lags at one day and over, sleep consolidation may play different roles during learning depending on the type of linguistic knowledge. Further research is needed to investigate the roles of consolidation during learning and then across the retention interval to the delayed posttest.

It is interesting to consider the extent to which the finding from study 1, which appeared to show that declarative memory and phonological short-term memory may interact with distributed practice to shift attention to different aspects of the input, support deficient processing accounts of the distributed practice effect (Challis, 1993; Greeno, 1967; Rundus, 1971; Zimmerman, 1975). In study 1, those with weaker phonological short-term memory and declarative memory appeared to remain focused on the most salient aspect of the artificial language (verbs) during massed practice and therefore actually did better in tests during training than those with stronger phonological short-term memory and declarative memory. This results may at first glance support deficient processing accounts, in which less attention is given, either consciously or subconsciously, during massed practice than distributed, thus resulting in weaker encoding and therefore worse retrieval (Challis, 1993). If weaker phonological short-term

memory means that during massed practice the item (verb) does not stay in working memory as long, then better encoding may result from subsequent presentations. However, deficient processing accounts do not explain the finding that those with stronger phonological short-term memory and declarative memory appeared to shift attention towards the adjectives and case markers as a result of distributed practice, resulting in higher test results during training. Indeed, if distributing practice appears to help those with stronger phonological short-term and declarative memory shift attention, and it is the attention that results in higher test scores, then this runs contrary to deficient processing accounts.

Type of Knowledge/Task being Learned

The two studies in this thesis shed some light on how distributed practice and different ISI/RI ratios interact with the type of knowledge or task being learned. One recent theory that has attempted to describe how declarative and procedural knowledge may interact with distributed practice is the skill retention theory by Kim et al. (2013). In their theory, declarative knowledge benefits from distributed practice to a much greater extent than procedural knowledge, due to the faster rate of forgetting in the former than the latter, although they do not appear to have taken into consideration offline gains during spacing in procedural tasks which may result in gains during the learning phase for distributed groups (e.g., Lee & Genovese, 1988). Based on a three-stage skill acquisition model (e.g., Anderson, 1982; DeKeyser, 2020; Fitts, 1964), comprised of acquiring (based on declarative memory), consolidating (procedural and declarative memory), and tuning (procedural memory) stages, they suggested that only the first two stages of skill acquisition benefit from distributed practice. However, it could be argued that learning a second language at the early stages under immersion conditions perhaps does not perfectly fit this three-stage acquisition model, with potentially much of the very early

stages remaining implicit. Kim et al. (2013) also suggested that the type of skill being trained includes varying profiles of the three stages. Therefore, more cognitive skills, which include a larger number of declarative components (e.g., Rohrer & Taylor, 2006), benefit more from spacing than more procedural skills (e.g., Vearrier et al., 2005).

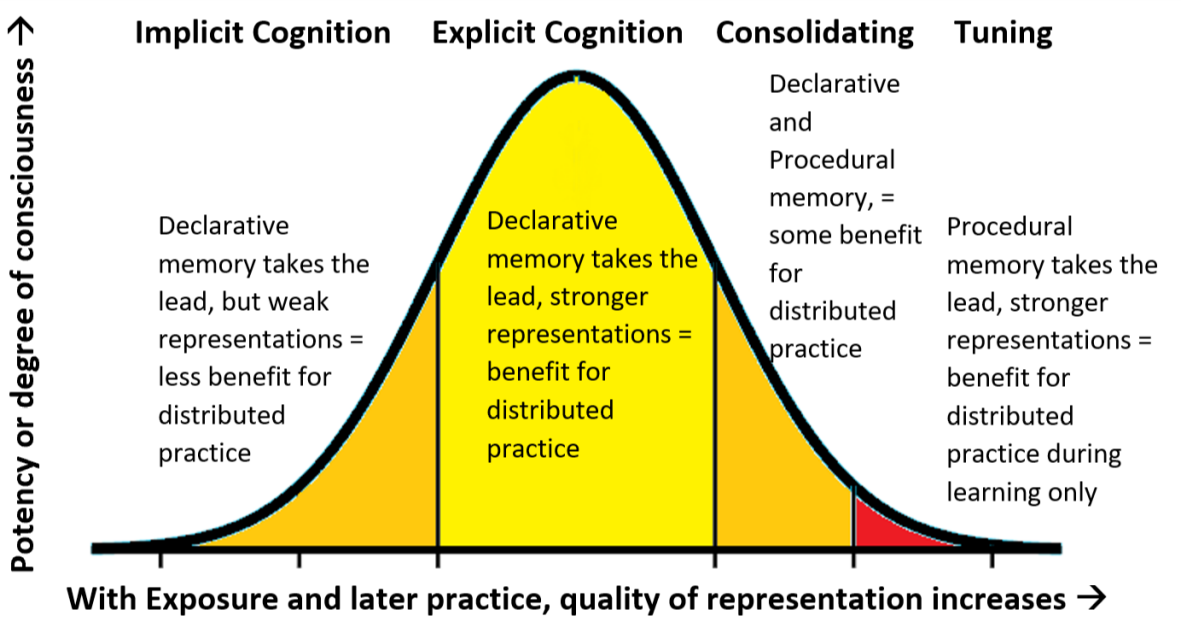
Another model, this time with regards to implicit, explicit and automated explicit knowledge, Cleeremans' (2007; 2011; see also Godfroid, 2022) radical plasticity thesis, offers an alternative perspective, focusing more on the early stages of second language learning under immersion conditions. In this model, initial implicit learning results in weak, low-quality representations but through further exposure, and as these representations become stronger and of a better quality, consciousness of that learning occurs, in what Godfroid (2022) calls a reverse interface.

6.6 An Alternative Model of Distributed Practice for L2 Learning

I now tentatively present a model of how declarative and procedural knowledge and memory systems might interact with distributed practice for L2 language learning (see Figure 17). This model adapts the skill retention theory of Kim et al. (2013) to include aspects of the radical plasticity thesis of Cleeremans (2007; 2011) and draws on findings from declarative and procedural SLA studies (Brill-Schuetz & Morgan-Short, 2014; Hamrick, 2015; Morgan-Short et al., 2014), as well as the findings from the two studies in this thesis. The model will also take on board the assertion, albeit with several caveats, from Kim et al. (2013) that procedural knowledge does not benefit from spacing as much as declarative knowledge due to slower forgetting rates for procedural memory. The first caveat is that the reason for a weaker distributed practice effect for procedural knowledge tasks may also be the result of the relative poor encoding, at least in the initial implicit cognition stage. Cleeremans (2007) suggests that in

this implicit cognition stage, there is a low depth of processing and that low-quality representations are formed. Therefore, it is possible that distributed exposure does not help as much (see Maddox, 2016 for a discussion of low strength of encoding), and that narrower lags are more optimal. The second caveat is that while distributed practice may not improve performance on procedural knowledge tasks in terms of retention at delayed posttests, it may improve performance during learning through offline consolidation (Lee & Genovese, 1988). It remains to be seen whether this is more so in the later automaticity (tuning) stage than in the early implicit cognition stage. Individual differences in procedural memory may then affect the amount of offline consolidation that occurs, as was hinted at in study 1.

Figure 17. An Updated Skill Retention Theory Model



Note. An update of the skill retention theory by Kim et al. (2013) combining aspects of Cleeremans' (2007, 2011) radical plasticity thesis.

Regarding the extent to which individual differences in declarative and procedural memory interact with different stages of this model, findings from SLA studies can help make

predictions. Declarative memory tends to predict success at early stages of acquisition and procedural memory at later stages (Hamrick, 2015; Morgan-Short et al., 2014). This suggests that in the very early implicit cognition stages of learning before the knowledge reaches the explicit cognition phase, distributed exposure may interact with declarative memory in terms of increasing abstraction, but to a lesser extent than during the explicit cognition stage when better encoding results in stronger mental representations. This supports the forgetting-as-abstraction theory of Vlach et al. (2012), in which commonalities between exemplars are remembered while differences are forgotten. It is also in line with the findings of the study by Vlach et al. (2008) in which there was a stronger distributed practice effect for straight memory task than a cross-situational word learning task.

Incidental and intentional learning conditions may influence which of the two memory systems take the primary role for learning (Brill-Schuetz & Morgan-Short, 2014). One difference between intentional and incidental learning conditions involves the amount of attentional resources that the participants pay to a particular to-be-learned item. Results of study 1 suggested that attention may affect which memory system takes the primary role. Declarative memory may take the primary role on items that are more attended to in the input, while procedural memory may take a more dominant role on items that are less attended to. Studies have shown that when attention is divided, declarative memory performance suffers (Fernandes & Moscovitch, 2013). Distributing practice, therefore, may be more beneficial to items that are attended to, either through choice, or through study design manipulation. An alternative, and potentially complimentary, reading of study 1 results suggests that the closer a learner is to the explicit cognition stage before the lag, that is, the better the quality of the representation and the stronger the encoding, the more likely it is that distributed exposure will be beneficial, and this in

turn may help shift attention towards other aspects of the to-be-learned items. The finding that declarative memory tends to play a stronger role for the learning of idiosyncratic items while procedural memory plays a stronger role for more pattern-like, sequential learning, and that more complex grammar may be better learned under incidental conditions (Brill-Schuetz & Morgan-Short, 2014) may also be explained partly by the amount of attention that is directed towards it. It follows that distributed practice may be less useful for more grammatical items.

In terms of lag and ISI/RI ratio, if an item is not encoded thoroughly in memory, as might be expected during the early stages of exposure under incidental conditions (implicit cognition stage), then retention will suffer, with long retention intervals resulting in catastrophic memory failure. Lags may only be of use in so far as they aid in the abstraction and the consolidation of knowledge. The benefit of this offline learning is likely to be in the region of hours and after a period of sleep.

The following final chapter concludes by outlining the theoretical, methodological and pedagogical contributions of this thesis. It then suggests three follow-up studies to further research distributed practice under incidental language learning conditions.

Chapter 7: Conclusion

The two studies in this thesis add to the current body of research into the temporal distribution of language learning under incidental learning conditions. Distributed practice was not significantly better than massed practice for the cross-situational learning paradigm in study 1 or for the incidental aspect of the form-meaning connections in study 2. However, elements of the results from both studies, including above chance results for a) certain aspects of the language (adjectives, study 1) or b) at particular lags (day 1 and 4, study 2), suggested that distribution of practice and exposure does alter the way that language is processed. The intentional aspect of the form-meaning connections in study 2 did produce a distributed practice effect. However, it did not produce one optimal lag but rather an optimal window. This suggests that more than one underlying mechanism of the distributed practice effect may be at work. Test items that required a generalisation of the rules did not produce results that were significantly different from those that had appeared in the exposure phase and therefore did not produce a narrower optimal lag. In study 1, individual differences in memory systems interacted with distribution of practice in ways that we did not predict. Results suggested that distributed practice, or at least differences in the length of lag, may encourage or force more of a reliance on individual differences. Declarative memory and phonological short-term memory appeared to interact with lag in terms of shifting attention, which in turn may alter the trajectory of learning. In study 2, surprisingly, declarative memory did not interact with lag. In chapter 6, I outlined a modified model of the skill retention theory by Kim et al. (2013) to include implicit cognition stages from the radical plasticity thesis by Cleeremans (2007, 2011).

This thesis has made contributions to theory, methodology and pedagogy. The theoretical contribution to the distributed practice effect involves adding further weight to

findings that L2 language learning under incidental learning conditions does not require the same optimal lag as intentional learning conditions (Verkoeijen et al., 2005), with massed practice possibly just as effective as distributed. This finding may help resolve some of the confusion around the mixed results of previous L2 grammar distributed practice effect studies (Bird, 2010; Kasprowicz et al., 2019; Miles, 2014; Rogers, 2015; Suzuki & DeKeyser, 2017a). Studies that involved more intentional learning of L2 grammar (Bird, 2010; Miles, 2014 and arguably Rogers, 2015) benefitted from lags and ISI/RI ratios within the 10-30% guidelines from Rohrer and Pashler (2007). The studies with more incidental learning conditions (Kasprowicz et al., 2019; Suzuki & DeKeyser, 2017a) demonstrated either benefits for the narrower lag or no advantage either way.

Another potentially significant theoretical contribution comes in the adaption of the skill retention theory of Kim et al. (2013). By combining the implicit cognition stage of the Cleereman's (2007, 2011) radical plasticity thesis with the abstraction by forgetting theory of Vlach et al. (2012), it provides a plausible explanation for how distributed practice interacts with the different stages of language acquisition, including the initial stages of learning under incidental, immersion conditions.

A further theoretical contribution of this thesis comes from a possible interpretation of the finding in study 1 that individual differences in declarative memory (or alternatively but just as intriguingly, visual working memory) together with phonological short-term memory may interact with lag to shift attention to other aspects of the language. In study 1, that shift was from verbs and basic word order to nouns and the noun phrase. Further research is needed to pin down both the reason for the three-way interaction and in the case of CVMT, what construct is helping shift attention.

In terms of contributions to the theory of cross-situational language learning, study 1 demonstrated that not only is learning more complex artificial languages that include nouns, verbs, adjectives and case markers possible under cross-situational learning conditions, but that learning is retained, and in some aspects, improved after 24 hours.

The methodological contribution of this thesis regards spacing and cross-situational learning study design and the terminology used in spacing studies. Many spacing SLA studies have two groups, one with a narrower ISI and one a wider (e.g., Bird, 2010; Kasprowicz et al., 2019; Suzuki & DeKeyser, 2017a), together with two RIs, often with two combinations of ISI/RI sitting within the optimal range of 10-30% ISI/RI suggested by Rohrer and Pashler (2007) and two groups outside that range. The results of study 2 suggest that, firstly, ISIs of 1-day should not be called "massed", as there are clear differences between 0-day and 1-day ISIs. Secondly, designing experiments with one optimal lag in mind, in which the two contrasting ISI/RI ratios are still within a window (e.g., 1-day ISI to 7-day ISI), may not produce results that reveal the interactions that may be present. Instead, perhaps including three groups, with a massed group acting as a base level, would provide a clearer picture of how distributed practice interacts with L2 language learning.

In terms of cross-situational learning methodological contributions, study 1 demonstrated that study designs which do not have a delayed posttest following a period of sleep may be under-capturing the learning that is taking place. Future cross-situational learning studies may benefit from including delayed posttests.

The pedagogical contribution of this thesis regards lessons from study 2 about optimal lags under intentional and incidental learning conditions. Learners, teachers and curriculum designers may wish to separate items learned under incidental conditions from those learned

under more deliberate, intentional conditions. Rote-learned vocabulary and explicitly taught rules of grammar that are to be remembered for a month or more may benefit from ISIs longer than a day. An optimal window of 24 hours to 30% ISI/RI ratio may help in the retention of idiosyncratic items. The more idiosyncratic the to-be-learned item (i.e., vocabulary), the closer the schedule can be to the ISI/RI ratios found by Cepeda et al. (2008), albeit with a wider optimal window rather than one peak. Results of the incidental aspect of the form-meaning connection in study 2, suggest, on the other hand, that for more pattern-based items (e.g., grammar) learned under more incidental learning conditions, it may be preferable to mass practice, or provide short lags (perhaps including sleep), that allow for the abstraction of rules and patterns. It should be noted that language classrooms rarely include significant chunks of input for language items to be learned incidentally. Instead, shorter awareness-raising tasks that guide learners to notice aspects of the language are more common, and these may benefit from either massed practice or short lags up to 24 hours. Input flooding in the form of extensive listening and reading to, say, films, music, novels or webpages is often left, understandably, to outside class. This thesis adds weight to the advice that learners should read and listen every day (Renandya & Jacobs, 2016). For low level learners who live in the L2 environment, the results of the cross-situational learning study 1 suggest that massing practice may be as effective as spaced practice. Neither of the studies focused on proceduralisation and automatisation of L2 language skills and systems, but the results of both studies add support to the skill retention theory of Kim et al. (2013), in which spacing may only be of partial assistance in the consolidation stage and of benefit in the tuning stage during practice. It may therefore be preferable to have shorter lags or even massed practice when providing learners with controlled and freer practice of target structures.

A likely future development for language learning apps may possibly involve integrating individual difference measures to provide individualised spacing schedules. Study 1 added to the evidence that individual differences in declarative memory, procedural memory and phonological short-term memory interact with distributed practice. The confirmation from study 1 that learning an artificial language under cross-situational learning conditions is possible and that learning is durable over 24 hours provides support for the next generation of language learning apps to include cross-situational learning conditions as a proxy for learning under immersion conditions.

Limitations to each of the two studies have been discussed in chapter 3, 4 and 5. However, here I will outline some of the broader limitations. Both studies were carried out with artificial language systems. While the artificial language in study 1 involved a grammar based on a real language (e.g., Japanese) and study 2 involved form-meaning connections of determiners that are found in real languages, it is yet to be determined whether the same results would occur with natural languages. Secondly, it is possible that different ISIs and RIs may have produced distributed practice effects under incidental learning conditions. For study 1, a wider ISI may have produced a benefit for distributed practice. For study 2, a shorter RI may have produced a significant advantage for the 1-day ISI compared to the massed group. The cost and logistics involved in including more ISI and RI combinations mean that the role of distributed practice under incidental learning conditions is necessarily limited to the combinations of ISI and RI that were included in the two studies.

Finally, I will briefly suggest three future research projects that may help build on the findings of the two studies in this thesis. The first potential study would investigate the role that distributed practice plays in shifting attention for those with strong declarative memory and

phonological short-term memory. It would also be useful to disentangle visual declarative memory from visual-spatial working memory by taking a measurement of the later with, for example, the computerised spatial span task (Woods et al., 2016). This could be carried out via eye-tracking methodology while participants carry out the alien cross-situational task. It would be useful to include a massed group and a longer lag of 1-day ISI and an RI of around 7 days.

The second suggestion involves partially replicating study 2, but with a) a shorter retention interval of, say, 14 days, and b) auditory cues followed by a response delay to ensure that online participants process the exposure sentences. It may also be beneficial to collect data face-to-face rather than online. It is possible that with these small modifications to the methodology, a distributed practice effect around 1-day ISI may be found.

The final suggestion would further investigate distributed exposure under incidental learning conditions by testing the claims by Kim et al. (2013) that items learned via procedural memory systems do not benefit from distributed exposure and my hypothesis that language learned through the procedural memory systems may benefit from distributed practice during the learning process via consolidation. This could be achieved by utilising the adaption by Batterink et al. (2014) of the experimental paradigm by Leung and Williams (2012), which captures measures of online implicit learning. In this possible study, lag groups of 0s (massed). 2 hours and 1 day would be used, with an RI of 7 days. Additionally, individual difference measures of procedural memory would be taken, to test whether it affects the amount of consolidation that occurs during the lag.

Reference List

Abend, O., Kwiatkowski, T., Smith, N. J., Goldwater, S., & Steedman, M. (2017). Bootstrapping language acquisition. *Cognition, 164, 116-143.*

Abraham, W. C. (2003). How long will long-term potentiation last? *Philos. Trans. R. Soc. Lond. B Biol. Sci*. 358, 735–744

Alario, F., & Cohen, L. (2004). Closed-class words in sentence production: Evidence from a modality-specific dissociation. *Cognitive Neuropsychology,, 21*(8), 787-819.

Ambridge, B., Theakston, A. L., Lieven, E. V., & Tomasello, M. (2006). The distributed learning effect for children's acquisition of an abstract syntactic construction. *Cognitive Development*, *21*(2), 174-193.

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review*, *89*(4), 369.

Andringa, S. (2020). The emergence of awareness in uninstructed L2 learning: A visual world eye tracking study. *Second Language Research*, *36*(3), 335-357.

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

Austin, S. D. M. (1921). A study in logical memory. *The American Journal of Psychology*, 370-403.

Baddeley, A. D., & Longman, D. J. A. (1978). The influence of length and frequency of training session on the rate of learning to type. *Ergonomics*, *21*(8), 627-635.

Baddeley, A., Papagno, C., & Vallar, G. (1988). When long-term learning depends on short-term storage. *Journal of Memory and Language, 27*(5), 586-595.

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*(5), 316-321.

Balota, D. A., Duchek, J. M., & Paullin, R. (1989). Age-related differences in the impact of spacing, lag, and retention interval. *Psychology and Aging*, *4*(1), 3.

Bannard, C., Lieven, E., & Tomasello, M. (2009). Modeling children's early grammatical knowledge. *Proceedings of the National Academy of Sciences*, *106*(41), 17284-17289.

Bates, E., & Goodman, J. (1997). On the inseparability of grammar and the lexicon: Evidence from acquisition, aphasia, and real-time processing. *Language and Cognitive Processes*, *12*, 507–584.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Bates, E., & MacWhinney, B. (1987). Competition, variation, and language learning. In MacWhinney, B. (Ed.), *Mechanisms of language acquisition* (pp. 157-193). Lawrence Erlbaum Associates.

Batterink, L. J., Oudiette, D., Reber, P. J., & Paller, K. A. (2014). Sleep facilitates learning a new linguistic rule. *Neuropsychologia*, *65*, 169-179.

Bell, M. C., Kawadri, N., Simone, P. M., & Wiseheart, M. (2014). Long-term memory, sleep, and the spacing effect. *Memory*, *22*(3), 276-283.

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

Berry, D. C., & Broadbent, D. E. (2014). Implicit learning in the control of complex systems. In P. Frensch & J. Funke (Eds.), *Complex problem-solving: The European perspective* (pp. 131–150). Psychology Press.

Bird, S. (2010). Effects of distributed practice on the acquisition of second language English syntax. *Applied Psycholinguistics*, *32*(2), 435-452.

Bjerrum, A. S., Eika, B., Charles, P., & Hilberg, O. (2016). Distributed practice. The more the merrier? A randomised bronchoscopy simulation study. *Medical Education Online*, *21*(1), 30517.

Bjork, R. A. (1994). Memory and metamemory considerations in the. *Metacognition: Knowing about Knowing*, *185*(7.2).

Bjork, R. A. (2018). Being suspicious of the sense of ease and undeterred by the sense of difficulty: Looking back at Schmidt and Bjork (1992). *Perspectives on Psychological Science*, *13*(2), 146-148.

Bliss T.V., & Collingridge, G.L. (1993). A synaptic model of memory: long-term potentiation in the hippocampus. *Nature*. 361(6407):31–9.

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.

Bock, K., & Levelt, W. (1994). Language production: Grammatical encoding. In Gernsbacher, M (Ed.*), Handbook of psycholinguistic*s (pp. 945-984). Academic Press.

Bod, R. (2009). From exemplar to grammar: A probabilistic analogy-based model of language learning. *Cognitive Science*, *33*(5), 752-793.

Brill-Schuetz, K. A., & Morgan-Short, K. (2014). The role of procedural memory in adult second language acquisition. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, 260-265.

Brooks, P. J., & Kempe, V. (2013). Individual differences in adult foreign language learning: The mediating effect of metalinguistic awareness. *Memory & Cognition*, *41*(2), 281-296.

Brown, R. M., & Robertson, E. M. (2007). Off-line processing: reciprocal interactions between declarative and procedural memories. *Journal of Neuroscience*, *27*(39), 10468-10475.

Bui, D. C., Maddox, G. B., Zou, F., & Hale, S. S. (2014). Examining the lag effect under incidental encoding: Contributions of semantic priming and reminding. *Quarterly Journal of Experimental Psychology*, *67*(11), 2134-2148.

Bui, G., Ahmadian, M. J., & Hunter, A.-M. (2019). Spacing effects on repeated L2 task performance. *System*, 81, 1–13.

Callan, D. E., & 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.

Camp, C. J., Foss, J. W., O'Hanlon, A. M., & Stevens, A. B. (1996). Memory interventions for persons with dementia. *Applied Cognitive Psychology*, *10*(3), 193-210.

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*, 369-378.

Carroll, J. B., & Sapon, S. M. (1959). *Modern language aptitude test*. Psychological Corporation.

Cecilio-Fernandes, D., Cnossen, F., Jaarsma, D. A., & Tio, R. A. (2018). Avoiding surgical skill decay: a systematic review on the spacing of training sessions. *Journal of Surgical Education*, *75*(2), 471-480.

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), 354–380.

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, and Cognition*, *19*(2), 389.

Chen, O., Castro-Alonso, J. C., Paas, F., & Sweller, J. (2018). Extending cognitive load theory to incorporate working memory resource depletion: evidence from the spacing effect. *Educational Psychology Review, 30*, 483-501.

Chen, W., Guo, X., Tang, J., Zhu, L., Yang, Z., & Dienes, Z. (2011). Unconscious structural knowledge of form–meaning connections. *Consciousness and Cognition*, *20*(4), 1751-1760.

Choi, S., & Gopnik, A. (1995). Early acquisition of verbs in Korean: A cross-linguistic study. *Journal of Child Language*, *22*(03), 497-529.

Christiansen, M. H. (2019). Implicit statistical learning: a tale of two literatures. *Topics in Cognitive Science*, *11*(3), 468-481.

Cleeremans, A. (2007). Consciousness: the radical plasticity thesis. *Progress in Brain Research*, *168*, 19-33.

Cleeremans, A. (2011). The radical plasticity thesis: how the brain learns to be conscious. *Frontiers in Psychology*, *2*, 86.

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.

Cornell, E. H. (1980). Distributed study facilitates infants' delayed recognition memory. *Memory & Cognition*, *8*(6), 539-542.

Coughlin, C. E., & Tremblay, A. (2013). Proficiency and working memory based explanations for nonnative speakers' sensitivity to agreement in sentence processing. *Applied Psycholinguistics*, *34*(3), 615-646.

Crowder, R. G. (1976/2014). *Principles of learning and memory: classic edition*. Psychology Press.

De Jong, N., & Perfetti, C. A. (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. *Language Learning*, 61(2), 533–568.

De Jong, N., & Tillman, P. (2018). Grammatical structures and oral fluency in immediate task repetition: Trigams across repeated performances. In M. Bygate (Ed.), *Language learning through task repetition* (pp. 43–73). John Benjamins.

DeKeyser, R. M. (2005). What makes learning second-language grammar difficult? A review of issues. *Language Learning*, *55*(S1), 1-25.

DeKeyser, R. (2012). Interactions between individual differences, treatments, and structures in SLA. *Language Learning*, *62*, 189-200.

DeKeyser, R. (2020). Skill acquisition theory. In *Theories in Second language acquisition* (pp. 83-104). Routledge.

Delaney, P. F., Verkoeijen, P. P., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. *Psychology of Learning and Motivation*, *53*, 63-147.

Dempster, F. N. (1987). Effects of variable encoding and spaced presentations on vocabulary learning. *Journal of Educational Psychology*, *79*(2), 162.

Desmottes, L., Meulemans, T., & Maillart, C. (2016). Later learning stages in procedural memory are impaired in children with specific language impairment. *Research in Developmental Disabilities*, *48*, 53-68.

Desmottes, L., Meulemans, T., Patinec, M. A., & Maillart, C. (2017). Distributed training enhances implicit sequence acquisition in children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, *60*(9), 2636-2647.

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*(5), 795.

Doughty, C. (1999). Cognitive underpinnings of focus on form. In P. Robinson (ed.), *Cognition and second language instruction*. Cambridge University Press, 206–257.

Dörnyei, Z. (2014). *The psychology of the language learner: Individual differences in second language acquisition*. Routledge.

Dudai, Y. (2004). The neurobiology of consolidations, or, how stable is the engram?. *Annu. Rev. Psychol.*, *55*, 51-86.

Durrant, S. J., Taylor, C., Cairney, S., & Lewis, P. A. (2011). Sleep-dependent consolidation of statistical learning. *Neuropsychologia*, *49*(5), 1322-1331.

Ebbinghaus, H. (1885/1964). Memory: A contribution to experimental psychology (H. A. Ruger, C. E. Bussenius Trans.). 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*(2), 365-414.

Ellenbogen, J. M., Hulbert, J. C., Stickgold, R., Dinges, D. F., & Thompson-Schill, S. L. (2006). Interfering with theories of sleep and memory: sleep, declarative memory, and associative interference. *Current Biology*, *16*(13), 1290-1294.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143-188.

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics*, *32*, 17-44.

Ellis, N. C., & Wulff, S. (2019). Cognitive approaches to second language acquisition. *The Cambridge handbook of language learning*, 41-61.

Erlam, R. (2005). Language aptitude and its relationship to instructional effectiveness in second language acquisition. *Language Teaching Research*, *9*(2), 147-171.

Evans, J. L., Saffran, J. R., & Robe-Torres, K. (2009). Statistical learning in children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, *52*(2), 321-335.

Fahl, J. T., Duvivier, R., Reinke, L., Pierie, J. P. E., & Schönrock-Adema, J. (2023). Towards best practice in developing motor skills: A systematic review on spacing in VR simulator-based psychomotor training for surgical novices. *BMC Medical Education*, *23*(1), 154.

Faretta-Stutenberg, M., & Morgan-Short, K. (2011). Learning without awareness reconsidered: A replication of Williams (2005). In G. Granena, J. Koeth, S. Lee-Ellis, A. Lukyanchenko, G. P. Botana, & E. Rhoades (Eds.), Selected Proceedings of the 2010 Second Language Research Forum: Reconsidering SLA Research, Dimensions, and Directions (pp. 18–28). Somerville, MA: Cascadilla Proceedings Project.

Fedzechkina, M., Newport, E. L., & Jaeger, T. F. (2016). Miniature artificial language learning as a complement to typological data. *The usage-based study of language learning and multilingualism.* Georgetown University Press.

Fernald, A., McRoberts, G., & Herrera, C. (1992). Prosodic features and early word recognition. In *8th International Conference on Infant Studies*, Miami, FL.

Fernandes, M. A., & Moscovitch, M. (2013). Divided attention and memory. In H. Pashler (Ed.), *Encyclopedia of the mind.* SAGE: Thousand Oaks.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 458.

Fitts, P. M. (1964). Perceptual-motor skill learning. In *Categories of human learning* (pp. 243-285). Academic Press.

Freudenthal, D., Pine, J. M., Aguado-Orea, J., & Gobet, F. (2007). Modeling the developmental patterning of finiteness marking in English, Dutch, German, and Spanish using MOSAIC. *Cognitive Science*, *31*(2), 311-341.

Friederici, A. D., Steinhauer, K., & Pfeifer, E. (2002). Brain signatures of artificial language processing: Evidence challenging the critical period hypothesis. *Proceedings of the National Academy of Sciences*, 99(1), 529-534.

Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: the paradox of statistical learning. *Trends in Cognitive Sciences*, *19*(3), 117-125.

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, *145*(12), 1128.

Frost, R. L., & Monaghan, P. (2016). Simultaneous segmentation and generalisation of non-adjacent dependencies from continuous speech. *Cognition*, *147*, 70-74.

Gathercole, S. E. (1995). Is nonword repetition a test of phonological memory or long-term knowledge? It all depends on the nonwords. *Memory & Cognition*, *23*(1), 83-94.

Gathercole, S. E., & Baddeley, A. D. (1996). *The children's test of nonword repetition*. Psychological Corporation.

Gentner D (1982) Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In: Kuczaj II S (ed.) Language development: Volume 2. Language, thought and culture. Lawrence Erlbaum, pp. 301–34.

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

Gerbier, E., Toppino, T. C., & Koenig, O. (2015). Optimising retention through multiple study opportunities over days: The benefit of an expanding schedule of repetitions. *Memory*, *23*(6), 943-954.

Gettinger, M., Bryant, N. D., & Fayne, H. R. (1982). Designing spelling instruction for learning-disabled children: An emphasis on unit size, distributed practice, and training for transfer. *The Journal of Special Education, 16*(4), 439-448.

Gleitman L (1990) The structural sources of verb meanings. Language Acquisition 1: 3–55.

Gleitman LR, Cassidy K, Nappa R, Papafragou A, and Trueswell JC (2005) Hard words. Language Learning and Development 1: 23–64.

Glenberg, A. M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Memory & Cognition*, *7*(2), 95-112.

Glenberg, A. M., & Lehmann, T. S. (1980). Spacing repetitions over 1 week. *Memory & Cognition*, *8*, 528-538.

Godfroid, A. (2022). Hypotheses about the interface between explicit and implicit knowledge in second language acquisition. In *The Routledge handbook of second language acquisition and psycholinguistics* (pp. 294-307). Routledge.

Goffman, L., & Gerken, L. (2020). An alternative to the procedural∼ declarative memory account of developmental language disorder. *Journal of Communication Disorders*, *83*, 105946.

Goldberg, A. (2006). Constructions at work. Oxford University Press.

Gómez, R. L., Bootzin, R. R., & Nadel, L. (2006). Naps promote abstraction in language-learning infants. *Psychological Science*, 17, 670-674.

Gómez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition,* 70(2), 109-135.

Goo, J., Granena, G., Yilmaz, Y., & Novella, M. (2015). Implicit and explicit instruction in L2 learning. *Implicit and Explicit Learning of Languages*, *48,* 443-482.

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

Greene, R. L. (1990). Spacing effects on implicit memory tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(6), 1004.

Greene, R. L., & Stillwell, A. M. (1995). Effects of encoding variability and spacing on frequency discrimination. *Journal of Memory and Language, 34*(4), 468-476.

Greeno, J. G. (1967). Paired-associate learning with short-term retention: Mathematical analysis and data regarding identification of parameters. *Journal of Mathematical Psychology*, *4*(3), 430-472.

Grey, S., Williams, J. N., & Rebuschat, P. (2015). Individual differences in incidental language learning: Phonological working memory, learning styles, and personality. *Learning and Individual Differences, 38*, 44-53.

Gupta, P. (2003). Examining the relationship between word learning, nonword repetition, and immediate serial recall in adults. *The Quarterly Journal of Experimental Psychology: Section A*, *56*(7), 1213-1236.

Guzman-Munoz, F. J. (2017). The advantage of mixing examples in inductive learning: A comparison of three hypotheses. *Educational Psychology*, *37*(4), 421-437.

Hagman, J. D. (1980). *Effects of training schedule and equipment variety on retention and transfer of maintenance skill*. Army Research Institute for the Behavioural and Social Sciences.

Hama, M., & Leow, R. P. (2010). Learning without awareness revisited: Extending Williams (2005). *Studies in Second Language Acquisition*, *32*(3), 465-491.

Hamrick, P. (2015). Declarative and procedural memory abilities as individual differences in incidental language learning. *Learning and Individual Differences*, *44*, 9-15.

Hawley, K. S., Cherry, K. E., Boudreaux, E. O., & Jackson, E. M. (2008). A comparison of adjusted spaced retrieval versus a uniform expanded retrieval schedule for learning a name–face association in older adults with probable Alzheimer's disease. *Journal of Clinical and Experimental Neuropsychology*, *30*(6), 639-649.

Hedenius, M., Persson, J., Tremblay, A., Adi-Japha, E., Veríssimo, J., Dye, C. D. & Ullman, M. T. (2011). Grammar predicts procedural learning and consolidation deficits in children with specific language impairment. *Research in Developmental Disabilities*, *32*(6), 2362-2375.

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

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

Housen, A., & Simoens, H. (2016). Introduction: Cognitive perspectives on difficulty and complexity in L2 acquisition. *Studies in Second Language Acquisition*, *38*(2), 163-175.

Hulstijn, J. H. (2005). Theoretical and empirical issues in the study of implicit and explicit second-language learning: Introduction. *Studies in Second Language Acquisition*, *27*(2), 129-140.

Hulstijn, J. H. (2015). Explaining phenomena of first and second language acquisition with the constructs of implicit and explicit learning. *Implicit and explicit learning of languages*, *48*, 25-46.

Hunt, R. H., & Aslin, R. N. (2001). Statistical learning in a serial reaction time task: Access to separable statistical cues by individual learners. *Journal of Experimental Psychology: General, 130*(4), 658.

Issa, B. I., & Morgan-Short, K. (2019). Effects of external and internal attentional manipulations on second language grammar development: An eye-tracking study. *Studies in Second Language Acquisition, 41(*2), 389–417

Jacoby, L. L. (1974). The role of mental contiguity in memory: Registration and retrieval effects. *Journal of Verbal Learning and Verbal Behavior*, *13*(5), 483-496.

Janiszewski, C., Noel, H., & Sawyer, A. G. (2003). A meta-analysis of the spacing effect in verbal learning: Implications for research on advertising repetition and consumer memory. *Journal of Consumer Research*, *30*(1), 138-149.

Jensen, T. D., & Freund, J. S. (1981). Persistence of the spacing effect in incidental free recall: The effect of external list comparisons and intertask correlations. *Bulletin of the Psychonomic Society*, *18*, 183-186.

Jones, G., & Rowland, C. F. (2017). Diversity not quantity in caregiver speech: Using computational modeling to isolate the effects of the quantity and the diversity of the input on vocabulary growth. *Cognitive Psychology*, 98, 1-21.

Kang, E. Y., Sok, S., & Han, Z. (2019). Thirty-five years of ISLA on form-focused instruction: A meta-analysis. *Language Teaching Research*, *23*(4), 428-453.

Kapa, L. L., Plante, E., & Doubleday, K. (2017). Applying an integrative framework of executive function to preschoolers with specific language impairment. *Journal of Speech, Language, and Hearing Research, 60*(8), 2170-2184.

Karpicke, J. D., & Roediger III, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*(2), 151-162.

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.

Kaufman, S. B., DeYoung, C. G., Gray, J. R., Jiménez, L., Brown, J., & Mackintosh, N. (2010). Implicit learning as an ability. *Cognition*, *116*(3), 321-340.

Kerz, E., Wiechmann, D., & Riedel, F. B. (2017). Implicit learning in the crowd: Investigating the role of awareness in the acquisition of L2 knowledge. *Studies in Second Language Acquisition*, *39*(4), 711-734.

Khoe, Y. H., Perfors, A., & Hendrickson, A. T. (2019). Modeling individual performance in cross-situational word learning. *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*.

Kidd, E. (2012). Implicit statistical learning is directly associated with the acquisition of syntax. *Developmental Psychology*, *48*(1), 171.

Kidd, E., & Arciuli, J. (2016). Individual differences in statistical learning predict children's comprehension of syntax. *Child Development*, *87*(1), 184-193.

Kim, A. S. N., Wong-Kee-You, A. M. B., Wiseheart, M., & Rosenbaum, R. S. (2019). The spacing effect stands up to big data. *Behavior Research Methods*, *51*, 1485-1497.

Kim, J. W., Ritter, F. E., & Koubek, R. J. (2013). An integrated theory for improved skill acquisition and retention in the three stages of learning. *Theoretical Issues in Ergonomics Science*, *14*(1), 22-37.

Kim, R., Seitz, A., Feenstra, H., & Shams, L. (2009). Testing assumptions of statistical learning: Is it long-term and implicit*? Neuroscience Letters, 461*(2), 145-149.

Kim, S. K., & Webb, S. (2022). The effects of spaced practice on second language learning: A meta-analysis. *Language Learning*, *72*(1), 269-319.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition, 83*(2), B35-B42.

Knabe, M. L., & Vlach, H. A. (2020). When are difficulties desirable for children? First steps toward a developmental and individual differences account of the spacing effect. *Journal of Applied Research in Memory and Cognition*, *9*(4), 447-454.

Kobayashi, M. (2022). The distributed practice effects of speaking task repetition. International *Journal of Applied Linguistics*, 32(1), 142–157.

Kóbor, A., Takács, Á., Kardos, Z., Janacsek, K., Horváth, K., Csépe, V., & Nemeth, D. (2018). ERPs differentiate the sensitivity to statistical probabilities and the learning of sequential structures during procedural learning. *Biological Psychology, 135*, 180-193.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"?. *Psychological Science*, *19*(6), 585-592.

Kornell, N., Castel, A. D., Eich, T. S., & Bjork, R. A. (2010). Spacing as the friend of both memory and induction in young and older adults. *Psychology and Aging*, *25*(2), 498.

Kornmeier, J., & Sosic-Vasic, Z. (2012). Parallels between spacing effects during behavioral and cellular learning. *Frontiers in Human Neuroscience*, *6*, 203.

Kornmeier, J., Spitzer, M., & Sosic-Vasic, Z. (2014). Very similar spacing-effect patterns in very different learning/practice domains. *PloS One*, *9*(3), e90656.

Koval, N. G. (2019). Testing the deficient processing account of the spacing effect in second language vocabulary learning: Evidence from eye tracking. *Applied Psycholinguistics*, *40*(5), 1103-1139.

Küpper-Tetzel, C. E. (2014). Strong effects on weak theoretical grounds: Understanding the distributed practice effect. *Z. f. Psychol*, *222*, 71-81.

Küpper-Tetzel, C. E., & Erdfelder, E. (2012). Encoding, maintenance, and retrieval processes in the lag effect: A multinomial processing tree analysis. *Memory*, *20*(1), 37-47.

Küpper-Tetzel, C. E., Erdfelder, E., & Dickhäuser, O. (2014). The lag effect in secondary school classrooms: Enhancing students' memory for vocabulary. *Instructional Science*, 42(3), 373–388.

Kwon, Y. H., Kwon, J. W., & Lee, M. H. (2015). Effectiveness of motor sequential learning according to practice schedules in healthy adults; distributed practice versus massed practice. *Journal of Physical Therapy Science*, *27*(3), 769-772.

Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, *8*(1), 33267.

Lambert, C., Kormos, J., & Minn, D. (2017). Task repetition and second language speech processing. *Studies in Second Language Acquisition*, *39*(1), 167-196.

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

Lattal, K. M. (1999). Trial and intertrial durations in Pavlovian conditioning: Issues of learning and performance. *Journal of Experimental Psychology: Animal Behavior Processes*, *25*(4), 433.

Lee, T. D., & Genovese, E. D. (1988). Distribution of practice in motor skill acquisition: Learning and performance effects reconsidered. *Research Quarterly for Exercise and Sport*, *59*(4), 277-287.

Lehmann, H., Sparks, F. T., Spanswick, S. C., Hadikin, C., McDonald, R. J., and Sutherland, R. J. (2009). Making context memories independent of the hippocampus. *Learning and Memory.* 16, 417–420.

Leow, R. P. (2018). Explicit learning and depth of processing in the instructed setting: Theory, research, and practice. *Studies in English Education, 23*(4), 769–801.

Leung, J. H., & Williams, J. N. (2012). Constraints on implicit learning of grammatical form-meaning connections. *Language Learning*, *62*(2), 634-662.

Leung, J. H., & Williams, J. N. (2014). Crosslinguistic differences in implicit language learning. *Studies in second language acquisition*, *36*(4), 733-755.

Levelt, W.J.M. (1989). Speaking: From intention to articulation. MIT Press.

Li, M. (2017). *Temporal distribution of practice and individual differences in the acquisition and retention of L2 Mandarin tonal word production* (Doctoral dissertation, University of Maryland, College Park).

Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. Language Learning, 60(2), 309–365.

Li, S. (2013). The interactions between the effects of implicit and explicit feedback and individual differences in language analytic ability and working memory. *The Modern Language Journal*, *97*(3), 634-654.

Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review, 21*(4), 861-883.

Lindsey, R. V., Shroyer, J. D., Pashler, H., & Mozer, M. C. (2014). Improving students' long-term knowledge retention through personalized review. *Psychological Science, 25*(3), 639-647.

Loewen, S., Erlam, R., & Ellis, R. (2009). The incidental acquisition of third person-s as implicit and explicit knowledge. In Elder, C. J., Ellis, R., Loewen, S., Elder, C. J., Erlam, R., Philp, J., &

Reinders, H. (2009). *Implicit and explicit knowledge in second language learning, testing and teaching*, 262-280. Multilingual Matters.

Logan, J. M., & Balota, D. A. (2008). Expanded vs. equal interval spaced retrieval practice: Exploring different schedules of spacing and retention interval in younger and older adults. *Aging, Neuropsychology, and Cognition*, *15*(3), 257-280.

Lum, J. A., Conti-Ramsden, G., Page, D., & Ullman, M. T. (2012). Working, declarative and procedural memory in specific language impairment. *Cortex*, *48*(9), 1138-1154.

Lum, J. A., Gelgic, C., & Conti-Ramsden, G. (2010). Procedural and declarative memory in children with and without specific language impairment. *International Journal of Language & Communication Disorders*, *45*(1), 96-107.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review, 101*(4), 676.

Macis, M., Sonbul, S., & Alharbi, R. (2021). The effect of spacing on incidental and deliberate learning of L2 collocations. *System*, *103*, 102649.

Mackey, A. (1999). Input, interaction, and second language development: An empirical study on question formation in ESL. *Studies in Second Language Acquisition*, *21*, 557–587.

Mackey, A., Adams, R., Stafford, C., & Winke, P. (2010). Exploring the relationship between modified output and working memory capacity. *Language Learning*, *60*(3), 501-533.

Mackey, A., & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 407–452). Oxford University Press.

Mackey, A., Kanganas, A. P., & Oliver, R. (2007). Task familiarity and interactional feedback in child ESL classrooms. *TESOL Quarterly*, *41*(2), 285-312.

Mackay, S., Morgan, P., Datta, V., Chang, A., & Darzi, A. (2002). Practice distribution in procedural skills training: a randomized controlled trial. *Surgical Endoscopy and Other Interventional Techniques*, *16*, 957-961.

Mackey, A., Philp, J., Egi, T., Fujii, A., & Tatsumi, T. (2002). Individual differences in working memory, noticing of interactional feedback, and L2 development. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 181–210). Benjamins.

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., 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*, 1376-1388.

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

Madsen, M. C. (1963). Distribution of practice and level of intelligence. *Psychological Reports*, *13*(1), 39-42.

Maki, R. H., & Hasher, L. (1975). Encoding variability: A role in immediate and long-term memory?. *The American Journal of Psychology*, 217-231.

Marcus, G. (1996). Why do children say "breaked"? *Current Directions in Psychological Science*, 5, 81–85.

Marshall, L., & Born, J. (2007). The contribution of sleep to hippocampus-dependent memory consolidation. *Trends in Cognitive Sciences*, *11*(10), 442-450.

Marshall, L., Mölle, M., Hallschmid, M., & Born, J. (2004). Transcranial direct current stimulation during sleep improves declarative memory. *Journal of Neuroscience*, *24*(44), 9985-9992.

Martin, K. I., & Ellis, N. C. (2012). The roles of phonological short-term memory and working memory in L2 grammar and vocabulary learning. *Studies in Second Language Acquisition*, *34*(3), 379-413.

Mauelshagen, J., Sherff, C. M., & Carew, T. J. (1998). Differential induction of long-term synaptic facilitation by spaced and massed applications of serotonin at sensory neuron synapses of Aplysia californica. *Learning & Memory*, *5*(3), 246-256.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101-B111.

McGregor, K. K., Sheng, L. I., & Smith, B. (2005). The precocious two-year-old: status of the lexicon and links to the grammar. *Journal of Child Language*, *32*(3), 563-585.

Meara, P. (2005). LLAMA language aptitude tests: The manual. Lognostics.

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, *108*(22), 9014-9019.

Menzel, R., Manz, G., Menzel, R., & Greggers, U. (2001). Massed and spaced learning in honeybees: the role of CS, US, the intertrial interval, and the test interval. *Learning & Memory*, *8*(4), 198-208.

Mettler, E., Burke, T., Massey, C. M., & Kellman, P. J. (2020). Comparing adaptive and random spacing schedules during learning to mastery criteria. In *CogSci.. Annual Conference of the Cognitive Science Society. Cognitive Science Society (US). Conference* (Vol. 2020, p. 773). NIH Public Access.

Meurers, D., De Kuthy, K., Nuxoll, F., Rudzewitz, B., & Ziai, R. (2019). Scaling up intervention studies to investigate real-life foreign language learning in school. *Annual Review of Applied Linguistics, 39*, 161-188.

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

Misyak, J. B., & Christiansen, M. H. (2012). Statistical learning and language: An individual differences study. *Language Learning*, *62*(1), 302-331.

Monaghan, P., & Mattock, K. (2012). Integrating constraints for learning word–referent mappings. *Cognition*, *123*(1), 133-143.

Monaghan, P., Mattock, K., Davies, R. A., & Smith, A. C. (2015). Gavagai is as Gavagai does: Learning nouns and verbs from cross-situational statistics. *Cognitive Science*, *39*(5), 1099-1112.

Monaghan, P., Ruiz, S., & Rebuschat, P. (2021). The role of feedback and instruction on the cross-situational learning of vocabulary and morphosyntax: Mixed effects models reveal local and global effects on acquisition. *Second Language Research*, *37*(2), 261-289.

Monaghan, P., Schoetensack, C., & Rebuschat, P. (2019). A single paradigm for implicit and statistical learning. *Topics in Cognitive Science*, *11*(3), 536-554.

Moranski, K., & Zalbidea, J. (2022). Context and generalizability in multisite L2 classroom research: The impact of deductive versus guided inductive instruction. *Language Learning*, *72*(S1), 41-82.

Morgan-Short, K., Faretta-Stutenberg, M., Brill-Schuetz, K. A., Carpenter, H., & Wong, P. C. (2014). Declarative and procedural memory as individual differences in second language acquisition. *Bilingualism: Language and Cognition*, *17*(1), 56-72.

Moulton, C. A. E., Dubrowski, A., MacRae, H., Graham, B., Grober, E., & Reznick, R. (2006). Teaching surgical skills: what kind of practice makes perfect?: A randomized, controlled trial. *Annals of surgery*, *244*(3), 400.

Mozer, M. C., Pashler, H., Cepeda, N., Lindsey, R. V., & Vul, E. (2009). Predicting the optimal spacing of study: A multiscale context model of memory. *Advances in Neural Information Processing Systems*, *22*.

Mumford, M. D., Costanza, D. P., Baughman, W. A., Threlfall, K., & Fleishman, E. A. (1994). Influence of abilities on performance during practice: Effects of massed and distributed practice. *Journal of Educational Psychology*, *86*(1), 134.

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., & Elgort, I. (2021). Effects of spacing on contextual vocabulary learning: Spacing facilitates the acquisition of explicit, but not tacit, vocabulary knowledge. *Second Language Research*, *37*(2), 233-260.

Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology*, *19*(1), 1-32.

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*(3), 417-528.

Paciorek, A., & Williams, J. N. (2015). Implicit learning of semantic preferences of verbs. *Studies in Second Language Acquisition*, *37*(2), 359-382.

Paik, J., & Ritter, F. E. (2016). Evaluating a range of learning schedules: hybrid training schedules may be as good as or better than distributed practice for some tasks. *Ergonomics*, *59*(2), 276-290.

Papagno, C., Valentine, T., & Baddeley, A. (1991). Phonological short-term memory and foreign-language vocabulary learning. *Journal of Memory and Language*, *30*(3), 331-347.

Paradis, M. (2009). *Declarative and procedural determinants of second languages* (Vol. 40). John Benjamins Publishing.

Peña, M., Bonatti, L. L., Nespor, M., & Mehler, J. (2002). Signal-driven computations in speech processing. *Science*, *298*(5593), 604-607.

Plihal, W., & Born, J. (1997). Effects of early and late nocturnal sleep on declarative and procedural memory. *Journal of Cognitive Neuroscience*, *9*(4), 534-547.

Quine, W. V. O. (1960). Word and object. Cambridge, MA: MIT Press.

Rakowska, M., Abdellahi, M. E., Bagrowska, P., Navarrete, M., & Lewis, P. A. (2021). Long term effects of cueing procedural memory reactivation during NREM sleep. *Neuroimage*, *244*, 118573.

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

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, *6*(6), 855-863.

Rebuschat, P. (Ed.). (2015). *Implicit and explicit learning of languages* (Vol. 48). John Benjamins Publishing Company.

Rebuschat, P., Hamrick, P., Riestenberg, K., Sachs, R., & Ziegler, N. (2015). Triangulating measures of awareness: A contribution to the debate on learning without awareness. *Studies in Second Language Acquisition*, *37*(2), 299-334.

Rebuschat, P., Hamrick, P., Sachs, R., Riestenberg, K., & Ziegler, N. (2013). Implicit and explicit knowledge of form-meaning connections: Evidence from subjective measures of awareness. In J.

M. Bergsleithner, S. N. Frota, & J. K. Yoshioka, (Eds.), *Noticing and second language acquisition: Studies in honor of Richard Schmidt* (pp. 255–275). University of Hawai'i, National Foreign Language Resource Center.

Rebuschat, P., & Monaghan, P. (2019). Editors' introduction: Aligning implicit learning and statistical learning: Two approaches, one phenomenon. *Topics in Cognitive Science*, *11*(3), 459-467.

Rebuschat, P., Monaghan, P., & Schoetensack, C. (2021). Learning vocabulary and grammar from cross-situational statistics. *Cognition*, *206*, 104475.

Renandya, W. A., & Jacobs, G. M. (2016). Extensive reading and listening in the L2 classroom. In W. A. Renandya, & Handoyo, P. (Eds.), *English language teaching today* (pp. 97-110). Springer International Publishing.

Reynolds, J. H., & Glaser, R. (1964). Effects of repetition and spaced review upon retention of a complex learning task. *Journal of Educational Psychology*, *55*(5), 297.

Riches, N. G., & Tomasello, M. Conti-Ramsden Gina (2005) Verb learning in children with SLI: Frequency and spacing effects. *Journal of Speech, Language, and Hearing Research, 48*, 1397 1411.

Robertson, E. M., Pascual-Leone, A., & Press, D. Z. (2004). Awareness modifies the skill-learning benefits of sleep. *Current Biology*, *14*(3), 208-212.

Robinson, P. (1997). Individual differences and the fundamental similarity of implicit and explicit adult second language learning. *Language Learning*, *47*(1), 45-99.

Robinson, P. (2005). Cognitive abilities, chunk-strength, and frequency effects in implicit artificial grammar and incidental L2 learning: Replications of Reber, Walkenfeld, and Hernstadt (1991) and Knowlton and Squire (1996) and their relevance for SLA. *Studies in Second Language Acquisition*, *27*(2), 235-268.

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

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

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

Rohrer, D., and Taylor, K. (2006). The Effects of Overlearning and Distributed Practise on the Retention of Mathematics Knowledge. *Applied Cognitive Psychology* 20 (9): 1209–1224.

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science, 35*, 481-498.

Rogers, J. (2015). Learning second language syntax under massed and distributed conditions. *TESOL Quarterly, 49*(4), 857-866.

Rogers, J. (2017). The spacing effect and its relevance to second language acquisition. *Applied Linguistics, 38*(6), 906-911.

Rogers, J. (2022). Spacing effects in task repetition research. *Language Learning*.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(6), 906-914.

Ruiz, S., Rebuschat, P., & Meurers, D. (2021). The effects of working memory and declarative memory on instructed second language vocabulary learning: Insights from intelligent CALL. *Language Teaching Research, 25*(4), 510-539.

Rumelhart, D. E., Greene Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. In D. E. Rumelhart & J. L. McClelland & the PDP Research Group

(Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1: Foundations* (pp. 110–146). MIT Press.

Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology*, 89(1), 63.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926-1928.

Sagarra, N., & Herschensohn, J. (2010). The role of proficiency and working memory in gender and number agreement processing in L1 and L2 Spanish. *Lingua*, *120*(8), 2022-2039.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11, 129–158.

Schmidt, R. (1995). Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning* (pp. 1– 63). University of Hawai'i Press.

Schmidt, R. (2010). Attention, awareness, and individual differences in language learning. In W. M. Chan, S. Chi, K. N. Cin, J. Istanto, M. Nagami, J. W. Sew, T. Suthiwan, & I. Walker (Eds.), *Proceedings of CLaSIC 2010*, Singapore, December 2–4 (pp. 721–737).National University of Singapore, Centre for Language Studies.

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.

Schoetensack, C. (2015). The role of working memory in vocabulary acquisition: A cross-situational word learning experiment. (Unpublished master's thesis), Lancaster University, Lancaster, United Kingdom.

Schönauer, M., Geisler, T., & Gais, S. (2014). Strengthening procedural memories by reactivation in sleep. *Journal of Cognitive Neuroscience*, *26*(1), 143-153.

Schvaneveldt, R. W., & Gómez, R. L. (1998). Attention and probabilistic sequence learning. *Psychological Research*, *61*(3), 175-190.

Scott, R. M., & Fisher, C. (2012). 2.5-year-olds use cross-situational consistency to learn verbs under referential uncertainty. *Cognition*, *122*(2), 163-180.

Seidenberg, M. S. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science*, *275*(5306), 1599-1603.

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*(3), 305-321.

Shafto, C. L., Conway, C. M., Field, S. L., & Houston, D. M. (2012). Visual sequence learning in infancy: Domain-general and domain-specific associations with language. *Infancy*, *17*(3), 247-271.

Shaughnessy, J. J. (1976). Persistence of the spacing effect in free recall under varying incidental learning conditions. *Memory & Cognition*, *4*, 369-377.

Shebilske, W. L., Goettl, B. P., Corrington, K., & Day, E. A. (1999). Interlesson spacing and task-related processing during complex skill acquisition. *Journal of Experimental Psychology: Applied*, *5*(4), 413.

Shintani, N., & Ellis, R. (2015). Does language analytical ability mediate the effect of written feedback on grammatical accuracy in second language writing?. *System*, *49*, 110-119.

Shoemaker, E., & Rast, R. (2013). Extracting words from the speech stream at first exposure. *Second Language Research, 29*(2), 165-183.

Siegelman, N., Bogaerts, L., Christiansen, M. H., & Frost, R. (2017). Towards a theory of individual differences in statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences. B, 372*(1711), 20160059.

Siegert, R. J., Taylor, K. D., Weatherall, M., & Abernethy, D. A. (2006). Is implicit sequence learning impaired in Parkinson's disease? A meta-analysis. *Neuropsychology, 20*(4), 490-495.

Simmons, A. L. (2012). Distributed practice and procedural memory consolidation in musicians' skill learning. *Journal of Research in Music Education, 59*(4), 357-368.

Simor, P., Zavecz, Z., Horváth, K., Éltető, N., Török, C., Pesthy, O., Gombos, F., Janacsek, K. and Nemeth, D (2019). Deconstructing procedural memory: Different learning trajectories and consolidation of sequence and statistical learning. *Frontiers in Psychology*, *9*, 2708.

Singh, L., Reznick, J.S., & Xuehua, L. (2012). Infant word segmentation and childhood vocabulary development: a longitudinal analysis. *Developmental Science, 15*(4), 482-495.

Sisti, H. M., Glass, A. L., & Shors, T. J. (2007). Neurogenesis and the spacing effect: learning over time enhances memory and the survival of new neurons. *Learning & Memory*, *14*(5), 368-375.

Slone, L. K., & Sandhofer, C. M. (2017). Consider the category: The effect of spacing depends on individual learning histories. *Journal of Experimental Child Psychology, 159*, 34-49.

Smith, C. D., & Scarf, D. (2017). Spacing repetitions over long timescales: a review and a reconsolidation explanation. *Frontiers in Psychology*, *8*, 962.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106, 1558–1568.

Smolak, E., McGregor, K. K., Arbisi-Kelm, T., & Eden, N. (2020). Sustained attention in developmental language disorder and its relation to working memory and language. *Journal of Speech, Language, and Hearing Research*, *63*(12), 4096-4108.

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning*, *60*(2), 263-308.

Speidel, G. E. (1993). Phonological short-term memory and individual differences in learning to speak: A bilingual case study. *First Language*, *13*(37), 69-91.

Sperber, R. D. (1974). Developmental changes in effects of spacing of trials in retardate discrimination learning and memory. *Journal of Experimental Psychology*, *103*(2), 204.

Soto, D., & Silvanto, J. (2014). Reappraising the relationship between working memory and conscious awareness. *Trends in Cognitive Sciences*, *18*(10), 520-525.

Staddon, J. E. R., Chelaru, I. M., & Higa, J. J. (2002). Habituation, memory and the brain: The dynamics of interval timing. *Behavioural Processes*, *57*(2-3), 71-88.

Stickgold, R., & Walker, M. P. (2005). Sleep and memory: the ongoing debate. *Sleep*, *28*(10), 1225-1227.

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. (2021). Optimizing fluency training for speaking skills transfer: Comparing the effects of blocked and interleaved task repetition. *Language Learning*, *71*(2), 285–325.

Suzuki, Y., & DeKeyser, R. M. (2017a). Effects of distributed practice on the proceduralization of morphology. *Language Teaching Research*.

Suzuki, Y., & Dekeyser, R. (2017b). Exploratory research on second language practice distribution: An aptitude× treatment interaction. *Applied Psycholinguistics*, *38*(1), 27-56.

Suzuki, Y., Eguchi, M., & De Jong, N. (2022). Does the reuse of constructions promote fluency development in task repetition? A usage-based perspective. TESOL Quarterly. Advance online publication.

Suzuki, Y., & Hanzawa, K. (2022). Massed task repetition is a double-edged sword for fluency development. *Studies in Second Language Acquisition, 44*(2), 536–561.

Suzuki, Y., Nakata, T., & Dekeyser, R. (2019). The desirable difficulty framework as a theoretical foundation for optimizing and researching second language practice. *The Modern Language Journal*, *103*(3), 713-720.

Suzuki, Y., Yokosawa, S., & Aline, D. (2022). The role of working memory in blocked and interleaved grammar practice: Proceduralization of L2 syntax. *Language Teaching Research, 26*(4), 671-695.

Tagarelli, K. M., Borges Mota, M.& Rebuschat, P. (2015). Working memory, learning conditions, and the acquisition of L2 syntax. In W. Zhisheng, M. Borges Mota, & A. McNeill (Eds.) *Working memory in second language acquisition and processing: Theory, research and commentary* (pp. 224 – 247). Multilingual Matters.

Tardif, T. (1996). Nouns are not always learned before verbs: Evidence from Mandarin speakers' early vocabularies. *Developmental Psychology, 32*(3), 492.

Thiessen, E. D., Kronstein, A. T., & Hufnagle, D. G. (2013). The extraction and integration framework: a two-process account of statistical learning. *Psychological Bulletin, 139*(4), 792.

Thiessen, E. D., & Saffran, J. R. (2007). Learning to learn: Infants' acquisition of stress-based strategies for word segmentation. *Language Learning and Development, 3*(1), 73-100.

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

Thompson, S. P., & Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development*, *3*(1), 1-42.

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, and Cognition*, *28*(3), 437.

Toppino, T. C., & DiGeorge, W. (1984). The spacing effect in free recall emerges with development. *Memory & Cognition*, *12*(2), 118-122.

Toppino, T. C., Fearnow-Kenney, M. D., Kiepert, M. H., & Teremula, A. C. (2009). The spacing effect in intentional and incidental free recall by children and adults: Limits on the automaticity hypothesis. *Memory & Cognition*, *37*(3), 316-325.

Trahan, D. E., & Larrabee, G. J. (1983). *Continuous visual memory test*. Psychological Assessment Resources.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, *66*(1), 126-156.

Tse, D., Langston, R. F., Kakeyama, M., Bethus, I., Spooner, P. A., Wood, E. R., et al. (2007). Schemas and memory consolidation. *Science* 316, 76–82.

Ullman, M. T. (2001). A neurocognitive perspective on language: The declarative/procedural model. *Nature Reviews Neuroscience*, *2*(10), 717-726.

Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition*, *92*(1), 231-270.

Ullman, M. T. (2016). The declarative/procedural model: A neurobiological model of language learning, knowledge and use. *The Neurobiology of Language*, 953-968.

Ullman, M. T., & Lovelett, J. T. (2018). Implications of the declarative/procedural model for improving second language learning: The role of memory enhancement techniques. *Second Language Research*, *34*(1), 39-65.

Ullman, M. T., & Pierpont, E. I. (2005). Specific language impairment is not specific to language: The procedural deficit hypothesis. *Cortex, 41*(3), 399-433.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*(3), 498-505.

VanPatten, B. (1996). *Input processing and grammar instruction in second language acquisition*. Greenwood Publishing Group.

VanPatten, B. (2002). Processing Instruction: An update. *Language Learning*, *52*(4), 755-803.

Vearrier, L. A., Langan, J., Shumway-Cook, A., & Woollacott, M. (2005). An intensive massed practice approach to retraining balance post-stroke. *Gait & Posture*, *22*(2), 154-163.

Verhagen, J., & Leseman, P. (2016). How do verbal short-term memory and working memory relate to the acquisition of vocabulary and grammar? A comparison between first and second language learners. *Journal of Experimental Child Psychology*, *141*, 65-82.

Verkoeijen, P. P., & Bouwmeester, S. (2008). Using latent class modeling to detect bimodality in spacing effect data. *Journal of Memory and Language*, *59*(4), 545-555.

Verkoeijen, P. P., Rikers, R. M., & Özsoy, B. (2008). Distributed rereading can hurt the spacing effect in text memory. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *22*(5), 685-695.

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

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

Vilberg, K. L., and Davachi, L. (2013). Perirhinal-hippocampal connectivity during reactivation is a marker for object-based memory consolidation. *Neuron* 79, 1232–1242.

Vlach, H. A., Bredemann, C. A., & Kraft, C. (2019). To mass or space? Young children do not possess adults' incorrect biases about spaced learning. *Journal of Experimental Child Psychology, 183*, 115-133.

Vlach, H. A., & DeBrock, C. A. (2017). Remember dax? Relations between children's cross-situational word learning, memory, and language abilities. *Journal of Memory and Language*, *93*, 217-230.

Vlach, H. A., Ankowski, A. A., & Sandhofer, C. M. (2012). At the same time or apart in time? The role of presentation timing and retrieval dynamics in generalization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(1), 246.

Vlach, H. A., & Johnson, S. P. (2013). Memory constraints on infants' cross-situational statistical learning. *Cognition*, *127*(3), 375-382.

Vlach, H. A., & Sandhofer, C. M. (2014). Retrieval dynamics and retention in cross-situational statistical word learning. *Cognitive Science*, *38*(4), 757-774.

Vlach, H. A., Sandhofer, C. M., & Kornell, N. (2008). The spacing effect in children's memory and category induction. *Cognition*, *109*(1), 163-167.

Vuong, L.C., Meyer, A.S. & Christiansen, M.H. (2016). Concurrent learning of adjacent and nonadjacent dependencies. *Language Learning*, 66, 8-30.

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, and Cognition*, *40*(1), 94.

Walker, N., Monaghan, P., Schoetensack, C., & Rebuschat, P. (2020). Distinctions in the acquisition of vocabulary and grammar: An individual differences approach. *Language Learning*, *70*(S2), 221-254.

Walker, N., Schoetensack, C., Monaghan, P., & Rebuschat, P. (2017). Simultaneous acquisition of vocabulary and grammar in an artificial language learning task. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 1307-1312). Cognitive Science Society.

Webb, S., & Chang, A. C. (2015). Second language vocabulary learning through extensive reading with audio support: How do frequency and distribution of occurrence affect learning?. *Language Teaching Research*, *19*(6), 667-686.

Wickelgren, W. A. (1972). Trace resistance and the decay of long-term memory. *Journal of Mathematical Psychology*. 9, 418–455.

Williams, J. N. (2005). Learning without awareness. *Studies in Second Language Acquisition*, *27*(2), 269-304.

Williams, J., & Evans, J. (1998). What kind of focus and on which forms? In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 139–155). Cambridge University Press.

Williams, J. N., & Kuribara, C. (2008). Comparing a nativist and emergentist approach to the initial stage of SLA: An investigation of Japanese scrambling. *Lingua*, 118(4), 522-553.

Williams, J. N., & Rebuschat, P. (2022). Implicit learning and second language acquisition: A cognitive psychology perspective. In *The Routledge handbook of second language acquisition and psycholinguistics* (pp. 281-293). Routledge.

Woods, D. L., Wyma, J. M., Herron, T. J., & Yund, E. W. (2016). An improved spatial span test of visuospatial memory. *Memory*, *24*(8), 1142-1155.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics: Research article. *Psychological Science*, 18, 414–420.

Yurovsky, D., & Frank, M. C. (2015). An integrative account of constraints on cross-situational learning. *Cognition, 145*, 53-62.

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

Appendices

Appendix A. Study 1: Pseudoword Lexicon

Adapted from Monaghan and Mattock (2012)

Bisyllabic words used content words (nouns, verbs, and adjectives): *barget, bimdah, chelad, dingep, fisslin, goorshell, haagle, jeelow, limeber, makkot, nellby, pakrid, rakken, sumbark*

Monosyllabic words used as function words (case markers): *tha, noo*

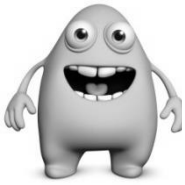
Appendix B. Study 1: Grammatical sentence patterns used in the cross-situational learning exposure task.

Word order	First phrase	Second phrase	Third phrase	N per block
SOV	NP_{subj} (Adj N $Case_{nom}$)	NP_{obj} (Adj N $Case_{acc}$)	VP (V)	2
	NP_{subj} (Adj N $Case_{nom}$)	NP_{obj} (N $Case_{acc}$)	VP (V)	3
	NP_{subj} (N $Case_{nom}$)	NP_{obj} (Adj N $Case_{acc}$)	VP (V)	1
	NP_{subj} (N $Case_{nom}$)	NP_{obj} (N $Case_{acc}$)	VP (V)	2
OSV	NP_{obj} (Adj N $Case_{acc}$)	NP_{subj} (Adj N $Case_{nom}$)	VP (V)	2
	NP_{obj} (Adj N $Case_{acc}$)	NP_{subj} (N $Case_{nom}$)	VP (V)	3
	NP_{obj} (N $Case_{acc}$)	NP_{subj} (Adj N $Case_{nom}$)	VP (V)	1
	NP_{obj} (N $Case_{acc}$)	NP_{subj} (N $Case_{nom}$)	VP (V)	2

Note: NP_{subj} = Subject Noun Phrase; NP_{obj} = Object Noun Phrase; VP = Verb Phrase; N = Noun; V = Verb; Adj = Adjective; $Case_{subj}$ = Case marker for subject; $Case_{obj}$ = Case marker for subject. N = Number of trials per block of 16 trials.

Appendix C. Study 1: Alien Characters Used in the Experiment.

Aliens 1 to 8 were used in exposure and testing trials and appeared in red or blue.

1. 
2. 
3. 
4. 
5. 
6. 
7. 
8. 

Appendix D. Study 1: Debriefing Questionnaire

1. During the different trials of this study, you saw two scenes and heard one sentence. Your task was to choose which scene the sentence referred to. How did you decide which scene the sentence described? Did you just guess throughout the experiment, or did you follow any particular strategies? If so, what strategies did you follow?

2. Do you think the way you made decision on the scenes changed throughout the experiment?

3. Did you feel you learned the names of the aliens?

4. Did you feel you learned the names of the colours?

5. Did you feel you learned the names of the actions?

6. Did you notice what type of word always preceded “tha”? If so, please write down the tope of word below. Also, please tell us when you noticed during the experiment (e.g., before the break etc.)? If you did not notice anything, please write down your best guess for what type of word precedes “tha”.

7. Did you notice what type of word always preceded “noo”? If so, please write down the tope of word below. Also, please tell us when you noticed during the experiment (e.g., before the break etc)? If you did not notice anything, please write down your best guess for what type of word precedes “noo”.

8. Do you think that “tha” had a particular grammatical function?

9. Do you think that “noo” had a particular grammatical function?

10. Did you notice any particular patterns or rules in the language while performing the task?

11. What do you think the aim of this study was?

Appendix E. Study 1: Summary of Reverse Helmert Contrasts for the Repeated Measures ANOVA for Nouns, Verbs, Adjectives, Case Markers and Word Order

	Nouns			Verbs			Adjectives			Case markers			Word Order		
			partial			partial			partial			partial			partial
	F	Sig	η2	F	Sig	η2	F	Sig	η2	F	Sig	η2	F	Sig	η2
Test 2 vs. test 1	7.10	.010	.10	3.07	.084	.057	0.27	.61	.004	0.96	.33	.015	8.60	.005	.12
Test 3 vs. previous	17.0	<.001	.21	1.06	.31	.027	0.38	.54	.006	0.65	.42	.010	7.49	.008	.11
Test 4 vs. previous	20.4	<.001	.25	3.93	.052	.059	8.4	.005	.12	0.31	.58	.005	7.61	.008	.11
Test 5 vs. previous	5.65	.021	.082	19.0	<.001	.23	0.29	.60	.005	2.07	.16	.032	4.14	.046	.062

Appendix F. Study 1: Descriptive Statistics and Summary of One-sample t-tests on Mean Scores for each Test Block.

			95% Confidence interval				
	M	*SD*	Lower	Higher	*t*	*df*	Sig. (2-tailed)
Noun test 1	.53	.18	.48	.57	1.14	63	.26
Noun test 2	.60	.19	.55	.65	4.16	63	<.001
Noun test 3	.68	.24	.62	.74	5.91	63	<.001
Noun test 4	.69	.23	.64	.75	6.82	63	<.001
Noun test 5	.68	.25	.62	.74	5.69	63	<.001
Verb test 1	.70	.25	.63	.76	6.16	63	<.001
Verb test 2	.76	.26	.69	.82	7.95	63	<.001
Verb test 3	.76	.28	.69	.83	7.49	63	<.001
Verb test 4	.79	.29	.72	.86	8.09	63	<.001
Verb test 5	.85	.22	.79	.90	12.8	63	<.001
Adjective test 1	.55	.26	.49	.61	1.58	63	.12
Adjective test 2	.53	.24	.47	.59	0.93	63	.36
Adjective test 3	.56	.26	.50	.63	1.95	63	.055
Adjective test 4	.64	.27	.58	.71	4.22	63	<.001
Adjective test 5	.59	.24	.53	.65	2.92	63	.005
Case marker test 1	.49	.17	.45	.54	-0.26	63	.80
Case marker test 2	.53	.16	.49	.56	1.32	63	.19
Case marker test 3	.53	.18	.49	.58	1.39	63	.17
Case marker test 4	.50	.22	.45	.56	0.086	63	.93
Case marker test 5	.54	.16	.50	.58	2.06	63	.043
Word order test 1	.76	.19	.72	.81	10.9	63	<.001
Word order test 2	.82	.20	.77	.87	13.0	63	<.001
Word order test 3	.84	.22	.79	.90	12.5	63	<.001
Word order test 4	.85	.22	.79	.90	12.6	63	<.001
Word order test 5	.85	.22	.79	.90	12.7	63	<.001

Appendix G. Study 1: Between-subject Effects of Repeated Measures ANOVA for Differences between Linguists vs. Non-linguistics, No Languages to Intermediate-level or Above vs. at least One Other Language to Intermediate-level or Above, No Previous Experience of Case-marked Language vs. Previous Experience of Case-marked Language, Degree vs. No Degree.

	Nouns			Verbs			Adjectives			Case markers			Word Order		
	F	Sig	partial η2	F	Sig	partial η2	F	Sig	partial η2	F	Sig	partial η2	F	Sig	partial η2
Linguist	0.35	.56	.014	0.38	.54	.016	1.83	.19	.071	0.49	.49	.009	1.57	.22	.062
Degree	0.44	.52	.018	0.30	.59	.013	0.86	.36	.034	0.05	.83	.001	0.02	.90	.001
Extra languages	0.41	.53	.017	0.51	.073	.13	0.95	.34	.038	0.40	.53	.008	2.23	.15	.085
Case-marked language	3.43	.069	.052	0.13	.72	.002	0.31	.58	.005	0.03	.86	.001	0.75	.39	.012

Appendix H. Study 1: Full Regression Table for Predicting Individual Difference Measures on Lexical Tests Divided by Massed and Distributed Groups

	Combined						Masssed						Distributed					
	Predicting	B	SEB	*β*	t	p	Predicting	B	SEB	*β*	T	p	Predicting	B	SEB	*β*	t	p
Nouns 1-4	0						0						0					
Noun test 1	0						0						0					
Noun test 2	zMLAT	0.055	0.024	0.283	2.306	0.025	0						zMLAT	0.110	0.043	0.428	2.551	0.016
Noun test 3	zMLAT	0.066	0.03	0.269	2.185	0.033	0						zSRT	0.082	0.038	0.369	2.140	0.041
Noun test 4	0						0						0					
Noun test 5	zSRT	0.088	0.030	0.349	2.904	0.005	0						zSRT	0.103	0.039	0.441	2.643	0.013
Nouns 1-5	0						0						0					
Verbs 1-4	0						0						zCVMT	-0.098	0.034	-0.436	-2.887	0.007
													zNRT	-0.067	0.027	-0.370	-2.451	0.021
Verb test 1	0																	

	Combined						Masssed						Distributed					
	Predicting	B	SEB	*β*	t	p	Predicting	B	SEB	*β*	T	p	Predicting	B	SEB	*β*	t	p
Verb test 2	0						0						0					
Verb test 3	zMLAT	0.082	0.034	0.292	2.385	0.020	0						0					
Verb test 4	0						0						0					
Verb test 5	0						0						0					
Verbs 1-5	0						0						0					
Adj test 1	0						0						0					
Adj test 2	0						zOSPAN	-0.081	0.038	-0.365	-2.149	0.040	zSRT	0.121	0.042	0.467	2.844	0.008
Adj test 3	0						0						0					
Adjective test 4	zSRT	0.111	0.032	0.402	3.428	0.001	zSRT	0.126	0.042	0.433	2.980	0.006	zSRT	0.134	0.048	0.463	2.809	0.009
							zNRT	0.138	0.048	0.418	2.877	0.007						
Adjective 1-4	zSRT	0.046	0.017	0.32	2.637	0.011	zNRT	0.058	0.028	0.352	2.058	0.048	zCVMT	0.070	0.025	0.428	2.804	0.009

													zSRT	0.059	0.024	0.379	2.483	0.019
	Combined						Masssed						Distributed					
	Predicting	B	SEB	*β*	t	p	Predicting	B	SEB	*β*	T	p	Predicting	B	SEB	*β*	t	p
Adjective test 5	zMLAT	0.081	0.029	0.339	2.812	0.007	0						zMLAT	0.121	0.043	0.466	2.834	0.008
Adjective 1-5	zSRT	0.047	0.017	0.343	2.848	0.006	zNRT	0.059	0.026	0.387	2.297	0.029	zMLAT	0.080	0.029	0.452	2.727	0.011
Marker test 1	0						0						0					
Marker test 2	0						0						zSRT	0.058	0.023	0.433	2.585	0.015
Marker test 3	0						0						zOSPAN	0.077	0.034	0.391	2.288	0.030
Marker test 4	0						0						zNRT	0.079	0.038	0.357	2.061	0.048
Marker 1-4	0						0						0					
Marker test 5	zMLAT	0.041	0.019	0.260	2.106	0.039	0						zMLAT	0.090	0.033	0.452	2.732	0.011
Markers 1-5	zSRT	0.025	0.012	0.262	2.122	0.038	0						zMLAT	0.047	0.022	0.362	2.093	0.045

	Combined						Masssed						Distributed					
	Predicting	B	SEB	*β*	t	p	Predicting	B	SEB	*β*	T	p	Predicting	B	SEB	*β*	t	p
Syntax 1-4	zMLAT	0.064	0.021	0.365	3.062	0.003	zMLAT	0.074	0.027	0.446	2.729	0.011	zCVMT	0.065	0.031	0.365	2.112	0.043
Syntax test 1	zCVMT	0.056	0.024	0.290	2.368	0.021	zMLAT	0.067	0.029	0.389	2.311	0.028	0					
Syntax test 2	zMLAT	0.075	0.024	0.376	3.173	0.002	zMLAT	0.093	0.034	0.448	2.742	0.010	0					
Syntax test 3	zCVMT	0.066	0.027	0.300	2.456	0.017	0						0					
Syntax test 4	zMLAT	0.078	0.027	0.351	2.931	0.005	zMLAT	0.076	0.037	0.349	2.042	0.050						
Syntax test 5	zMLAT	0.063	0.027	0.286	2.331	0.023	zMLAT	0.083	0.038	0.370	2.183	0.037	0					
Syntax 1-5	zMLAT	0.059	0.021	0.342	2.846	0.006	zMLAT	0.075	0.026	0.457	2.816	0.009	0					

Appendix I. Study 1 and 2: Ethical Approval

Study 1: Participant Information Sheet

For further information about how Lancaster University processes personal data for research purposes and your data rights please visit our webpage: www.lancaster.ac.uk/research/data-protection

I am a PhD student at Lancaster University and I would like to invite you to take part in a research study about individual differences in language processing.

Please take time to read the following information carefully before you decide whether or not you wish to take part.

What is the study about and why have I been invited?

This is an experiment about individual differences in language processing. I am also interested in whether this changes over time. I would be very grateful if you would agree to take part in this study. You have been invited to partiipate in the study because you are a native speaker of English.

What will I be asked to do if I take part?

If you decided to take part, this would involve the following:

The study involves completing a series of tasks on the computer as well as a short debriefing questionnaire. In total, the study takes between 1 and 2 hours to complete (with breaks). You will then be asked to return 24 hours later for another half hour.

You will be randomly assigned to one of two groups:

Group 1: approximately 1 hour on day 1, half an hour on day 2.

Group 2: approximately 1 hour 45 minutes on day 1, half an hour on day 2.

You will be paid eiher £20 or £28 depending on which of the two groups you are randomly assigned. This reflects the differing amount of time you will have to commit to the project. In order to be paid, you will need to attend both sessions. Therefore, before you agree to take part in this experiment please ensure that you are available on both days.

What are the possible benefits from taking part?

If you take part in this study, your insights will contribute to our understanding of language processing and language practice. You may alos gain some insights into the way that you learn languages.

Do I have to take part?

No. It's completely up to you to decide whether or not you take part. Your participation is voluntary.

If you are a student here and you decide not to take part in this study, this will not affect your studies and the way you are assessed on your course.

What if I change my mind?

If you change your mind, you are free to withdraw at any time during your participation in this study. If you want to withdraw, please let me know, and I will extract any ideas or information (=data) you contributed to the study and destroy them. However, it is difficult and often impossible to take out data from one specific participant when this has already been anonymised or pooled together with other people's data. Therefore, you can only withdraw up to 6 weeks after taking part in the study.

What are the possible disadvantages and risks of taking part?

It is unlikely that there will be any major disadvantages to taking part. However, you will need to invest between 90 – 145 minutes following the specified schedule.

Will my data be identifiable?

After taking part in the experiment, only my supervisor and I, the researcher conducting this study, will have access to the data that provide me.

I will keep all personal information about you (e.g., your name and other information about you that can identify you) confidential, that is I will not share it with others. I will remove any personal information from the written record of your contribution.

How will we use the information you have shared with us and what will happen to the results of the research study?

I will use the information you have shared with me only in the following ways:
I will use it for research purposes only. This will include my PhD thesis and journal articles. I may also present the results of my study at academic conferences.

When writing up the findings from this study, I would like to reproduce some of the views and ideas you shared with me. I will only use anonymised quotes (e.g., from the debriefing questionnaire with you), so that although I will use your exact words, you cannot be identified in our publications.

How my data will be stored
Your data will be stored in encrypted files (that is no-one other than me, the researcher will be able to access them) and on password-protected computers. I will store hard copies of any data

securely in locked cabinets in my office. I will keep data that can identify you separately from non-personal information (e.g.,your views on a specific topic).In accordance with University guidelines, I will keep the data securely for a minimum of ten years.

What if I have a question or concern?

If you have any queries or if you are unhappy with anything that happens concerning your participation in the study, please contact me, Neil Walker (**n.walker@lancaster.ac.uk**, 01772 893151), or my supervisor, Patrick Rebuschat (**p.rebuschat@lancaster.ac.uk**)

If you have any concerns or complaints that you wish to discuss with a person who is not directly involved in the research, you can also contact:

Uta Papen, u.papen@lancaster.ac.uk, **Department: Linguistics and English Language, County South Building, Lancaster University.**
Tel: 01524 593245

Thank you for considering your participation in this project.

Study 1: Consent Form

UNIVERSITY OF LANCASTER

Department of Linguistics and English Language

Consent Form

Project title: Individual variation in adult language processing

1. I have read and had explained to me by Neil Walker the Information Sheet relating to this project.

2. I have had explained to me the purposes of the project and what will be required of me, and any questions have been answered to my satisfaction. I agree to the arrangements described in the Information Sheet in so far as they relate to my participation.

3. I understand that my participation is entirely voluntary and that I have the right to withdraw from the project any time.

4. I have received a copy of this Consent Form and of the accompanying Information Sheet.

Name:

Signed:

Date:

Study 1: Ethical Approval

3rd March 2016

Neil Walker
School of Journalism, Language and Communication
University of Central Lancashire

Dear Neil,

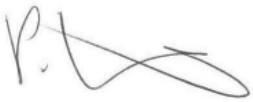
Re: BAHSS Ethics Committee Application
Unique reference Number: BAHSS 333

The BAHSS ethics committee has granted approval of your proposal application 'Age effects in second language acquisition: An artificial language approach (Adult participants only)'. Approval is granted up to the end of project date* or for 5 years from the date of this letter, whichever is the longer. It is your responsibility to ensure that:

- the project is carried out in line with the information provided in the forms you have submitted
- you regularly re-consider the ethical issues that may be raised in generating and analysing your data
- any proposed amendments/changes to the project are raised with, and approved, by Committee
- you notify roffice@uclan.ac.uk if the end date changes or the project does not start
- serious adverse events that occur from the project are reported to Committee
- a closure report is submitted to complete the ethics governance procedures (Existing paperwork can be used for this purposes e.g. funder's end of grant report; abstract for student award or NRES final report. If none of these are available use e-Ethics Closure Report Proforma).

Please also note that it is the responsibility of the applicant to ensure that the ethics committee that has already approved this application is either run under the auspices of the National Research Ethics Service or is a fully constituted ethics committee, including at least one member independent of the organisation or professional group.

Yours sincerely,



Peter Lucas
Chair
BAHSS Ethics Committee

* for research degree students this will be the final lapse date

Study 2: Ethical Approval

From: FASS and LUMS Research Ethics <fass.lumsethics@lancaster.ac.uk>
Sent: 05 November 2019 14:49
To: Walker, Neil (Student) <n.walker@lancaster.ac.uk>
Cc: Rebuschat, Patrick <p.rebuschat@lancaster.ac.uk>
Subject: Ethics approval (reference FL19008) please quote this reference in all correspondence about this project

Dear Neil

Thank you for submitting your application and additional information for *Spaced Practice under intentional and incidental learning conditions.* The information you provided has been reviewed by members of the Faculty of Arts and Social Sciences and Lancaster Management School Research Ethics Committee and I can confirm that approval has been granted for this project.

As principal investigator your responsibilities include:

- ensuring that (where applicable) all the necessary legal and regulatory requirements in order to conduct the research are met, and the necessary licenses and approvals have been obtained;
- reporting any ethics-related issues that occur during the course of the research or arising from the research (e.g.,unforeseen ethical issues, complaints about the conduct of the research, adverse reactions such as extreme distress) to the Research Ethics Officer;
- submitting details of proposed substantive amendments to the protocol to the Research Ethics Officer for approval.

Please do not hesitate to contact me if you require further information about this.

Kind regards,

Debbie

Debbie Knight

Secretary, FASS-LUMS Research Ethics Committee fass.lumsethics@lancaster.ac.uk I Phone (01524) 592605| A04 Bailrigg House, Lancaster University, LA1 4YT I Web: FASS & LUMS Research Ethics Guidance & Application form

Lancaster University

www.lancaster.ac.uk/50

Lancaster University

Study 2: Participant information sheet

For further information about how Lancaster University processes personal data for research purposes and your data rights please visit our webpage: www.lancaster.ac.uk/research/data-protection

I am a PhD student at Lancaster University and I would like to invite you to take part in a research study about the way we formulate and process sentences in different languages over time.

Please take time to read the following information carefully before you decide whether or not you wish to take part.

What is the study about and why have I been invited?

This is an experiment about how people formulate sentences in different languages. Different languages have different ways of saying things, which raises the question of whether this forces people to think slightly differently in order to speak and understand in different languages. I am also interested in whether this changes over time. I would be very grateful if you would agree to take part in this study.

What will I be asked to do if I take part?

If you decided to take part, this would involve the following:

You will spend approximately an hour (60-70 mins) spread over four sittings (each ten – twenty minutes long) completing an online task which involves reading sentences and making decisions.

You will need to complete all four sittings at the specified times, which will be clearly given to you before you agree to take part in this experiment together with reminder emails 24 hours before each session.

For example:

Session 1	Session 2	Session 3	Session 4
April 13th	April 20th	April 27th	May 2nd
15 mins	15 mins	20mins	20mins

In order to be paid £10, you will need to complete all four sections of the experiment in the timeframe specified. Therefore, before you agree to take part in this experiment please ensure that you are available on the given days. I recommend that you put the dates into your calendar with reminder notices.

In the fourth sitting you will also complete a short debriefing questionnaire that will explore your experience of carrying out the prior task.

What are the possible benefits from taking part?

If you take part in this study, your insights will contribute to our understanding of language processing and practice.

Do I have to take part?

No. It's completely up to you to decide whether or not you take part. Your participation is voluntary.

If you are a student here and you decide not to take part in this study, this will not affect your studies and the way you are assessed on your course.

What if I change my mind?

If you change your mind, you are free to withdraw at any time during your participation in this study. If you want to withdraw, please let me know, and I will extract any ideas or information (=data) you contributed to the study and destroy them. However, it is difficult and often impossible to take out data from one specific participant when this has already been anonymised or pooled together with other people's data. Therefore, you can only withdraw up to 6 weeks after taking part in the study.

What are the possible disadvantages and risks of taking part?

It is unlikely that there will be any major disadvantages to taking part. However, you will need to invest between 60-70 minutes over 4 sessions following the specified schedule.

Will my data be identifiable?

After taking part in the experiment, only my supervisor and I, the researcher conducting this study, will have access to the data.

I will keep all personal information about you (e.g., your name and other information about you that can identify you) confidential, that is I will not share it with others. I will remove any personal information from the written record of your contribution.

How will we use the information you have shared with us and what will happen to the results of the research study?

I will use the information you have shared with me only in the following ways:
I will use it for research purposes only. This will include my PhD thesis and journal articles. I may also present the results of my study at academic conferences.

When writing up the findings from this study, I would like to reproduce some of the views and ideas you shared with me. I will only use anonymised quotes (e.g.,from the debriefing questionnaire with you), so that although I will use your exact words, you cannot be identified in our publications.

How my data will be stored

Your data will be stored in encrypted files (that is no-one other than me, the researcher will be able to access them) and on password-protected computers. I will store hard copies of any data securely in locked cabinets in my office. I will keep data that can identify you separately from non-personal information (e.g.,your views on a specific topic).In accordance with University guidelines, I will keep the data securely for a minimum of ten years.

What if I have a question or concern?

If you have any queries or if you are unhappy with anything that happens concerning your participation in the study, please contact me, Neil Walker (**n.walker@lancaster.ac.uk**, 01772 893151), or my supervisor, Patrick Rebuschat (**p.rebuschat@lancaster.ac.uk**)

If you have any concerns or complaints that you wish to discuss with a person who is not directly involved in the research, you can also contact:

Uta Papen, u.papen@lancaster.ac.uk, **Department: Linguistics and English Language, County South Building, Lancaster University.**
Tel: 01524 593245

Thank you for considering your participation in this project.

Study 2: Consent Form

Project Title: Language Formulation

Name of Researchers: Neil Walker

Email: n.walker@lancaster.ac.uk

Please tick each box

1. I confirm that I have read and understand the information sheet for the above study. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily	☐
2. I understand that my participation is voluntary and that I am free to withdraw at any time during my participation in this study and within 4 weeks after I took part in the study, without giving any reason. If I withdraw within 4 weeks of taking part in the study my data will be removed. If I am involved in focus groups and then withdraw my data will remain part of the study.	☐
3. I understand that any information given by me may be used in future reports, academic articles, publications or presentations by the researcher/s, but my personal information will not be included and I will not be identifiable.	☐
4. I understand that my name/my organisation's name will not appear in any reports, articles or presentation without my consent.	☐
5. I understand that data will be kept according to University guidelines for a minimum of 10 years after the end of the study.	☐
6. I agree to take part in the above study.	☐

_____ _____ _____

Name of Participant Date Signature

I confirm that the participant was given an opportunity to ask questions about the study, and all the questions asked by the participant have been answered correctly and to the best of my ability. I confirm that the individual has not been coerced into giving consent, and the consent has been given freely and voluntarily.

Signature of Researcher /person taking the consent_____ Date _____ Day/month/year

One copy of this form will be given to the participant and the original kept in the files of the researcher at Lancaster University

Appendix J. Study 2: Comparison of the Study Designs for Williams (2005), Hama and Leow (2010), Faretta-Stuttenberg and Morgan-Short (2011) and Rebuschat et al. (2013)

Authors	Number of participants	Mean age	% female	From?	Non-native	Items
Williams, (2005)	41	24.9	68%	Undergraduate or graduate students of Cambridge 34% language or linguistics	34%	12 nouns for each class -i.e. 24 in total 8 used with both determiners- but with either singular or plural. 4 nouns with only one determiner in singular and plural
Hama & Leow (2010)	34 After excluding 42 for not completing task And 11 Non-animacy based strategy	-	-	- None from linguistics background	-	Williams (2005) + 16 new noun phrases (8 with training and 8 with test) So 16 for each class in total 8 used with both determiners and either singular or plural. 8 only with one determiner in singular and plural
Faretta-Stutenberg & Morgan-Short (2011)	30	18.93	67%	Undergraduates Psychology	0% 50% had another L1 as well (11 L1s encoded grammatical gender) 14 L2 at intermediate or above Average of .80 No language-related students	Williams (2005)
Rebuschat, Hamrick, Sachs, Riestenberg & Ziegler (2013)	30 15 in trained and 15 in control	20	60%	Georgetown. 11 from linguistics	0%, 12 had another L1; 27 had studied another language	Same as Williams, 2005 experiment 2, i.e. 12 nouns in each class Each used with singular and plural, but with same determiner

Author(s)	Test items	Modality	Number of sets	Flip determiners	Background questionnaire	Memory test	Time	ID	Vocabulary pre-training
Williams, (2005)	Singular and plural form of withheld items with other determiner And 10 trained items	Oral training Written/visual test	2 Each set included all nouns, each noun had either a different determiner or for generalised items singular or plural	Yes	Yes	Yes	1hr	PSTM	Cards with colours Comprehension task via computer
Hama & Leow (2010)	MC: 8 trained and 8 new Production: 7 new, 8 trained	Oral training and test		No	Yes	Yes	-		Comprehension task – 12 times per word in random order + think aloud
Faretta-Stutenberg & Morgan-Short (2011)	2 trained, 8 generalised, 8 trained, 8 generalised	Oral training Written/visual test	2	No	Yes	Yes, 10 from training	90mins		Cards with colours Computer flashcards for production and comprehension 12 times
Rebuschat, Hamrick, Sachs, Riestenberg & Ziegler (2013)	36 completely new sentences Trained (12), partially trained (12 – different distance determiner), new (12 – new NP)	Written/visual training Written/visual test	2 sets, each set had 6 NPs of each type. They only differed in terms of singular and plural	No	Yes	No	-		Powerpoint Flashcards. Comprehension test, 12 repetitions of each, including saying it aloud. Feedback given Given test -offered opportunity to repeat

Author(s)	Task	Training	Rest	Test	Test instructions	Test procedure
Williams, (2005)	Listen Press near or far Repeat aloud Form a mental image Non-native – only noun phrase e.g.,gi lion Told how many sentences. Three repetitions of each. Told would be a memory test but not necessary to remember exact wording just general situation	6 blocks of 24 trials, set 1, set 2, set 1, set 2, set 1, set 2	30 secs between blocks	2-way MC (gi clock or ro clock)	More familiar, better, or more appropriate	2 trained items, 8 generalised – once -equal number of singular and plural with each determiner 8 trained 8 generalised - opposite singular or plural and in different contexts
Hama & Leow (2010)	Listen Repeat exactly Choose near or far and give reasoning (think aloud) Form a mental image Told would be a memory test but not necessary to remember exact wording just general situation	Set 1, Set 2, Set 1, Set 2, Set 1, Set 2	-	4-way MC + production task Listen to sentence – determiner has beep a) gi clock, b) ro clock c) Ul clock d) ne clock Production: sentence with beep, asked to think aloud (before MC task)	-	Same sentences as Williams (2005) but modified to include distance information. 2 trained, 8 new, 8 trained, 8 new
Faretta-Stutenberg & Morgan-Short (2011)	Listen Press near or far Repeat aloud Form a mental image Told would be a memory test but not exact wording just general situation	6 blocks of 24 trials, set 1, set 2, set 1, set 2, set 1, set 2	-	2-way MC (gi clock or ro clock)	More familiar, better, or more appropriate	-
Rebuschat, Hamrick, Sachs, Riestenberg & Ziegler (2013)	Read Press near or far Repeat noun phrase Form a mental image 4 practice sentences Told that forming mental image was an important part of the study	Set 1, set 2 , set 1, set 2, set 1, set 2	-	2-way MC (gi clock or ro clock)	More familiar, better, or more appropriate	Same as Williams (2005)

	Checking awareness	Awareness results
Williams, (2005)	Asked about rule (animacy or moving/not moving) When did you become aware? If no rule, tell them there is a rule. 10 trained, 8 generalised, 8 generalised Asked for rule and when found. Or given rule and asked if considered	8/41 after test 1, 11 after rule search 6 during training, 2 during test 91.4% for generalised items 6 described wrong rule 27 familiarity or intuition All 33 not considered rule All 33 above chance even for generalised 6 with wrong rule only 49% for generalised But 68% for trained Rule-search test 2 11 aware – sig above chance 22 unaware – also sig above chance
Hama & Leow (2010)	Think aloud protocols (understanding, noticing, nothing) + post test questionnaire similar to Williams (2005)	9 aware Think aloud: none during training, 3 gave right rule and 5 described partial rule during testing Aware: MC: 79% for trained and 81% new; Production: 92% trained and 96% new Unaware: Trained above chance for MC: animacy =53% for trained and 49% new; distance = 83% trained, 91% new; Main effect for distance with trained and new No effect for animacy; Same results for production test
Faretta-Stutenberg & Morgan-Short (2011)	Debriefing questionnaire. What criteria they used to make their decision in the test; then, how and when they knew to use each of the 4 determiners. If no rule, ask if they found a rule. If no rule, then rule search. 10 trained, 15 generalised; then asked for rule; asked about animacy Used Williams, 2005) aware or unaware, and Hama and Leow, (2010) unaware, noticing, understanding	9/30 Aware: above chance for trained (76%) but not generalised (58.3%)on test 1. Unaware: same (62.3%, 53.9%) 2nd protocol: 4 understanding (above chance for trained (92.5%) and 2nd generalised (75%); 13 noticing (not above chance for trained, 60.7%; generalised, 56.7%) Unaware: (above chance for T, 64% but not G, 50.5%); 2nd test phase, 2 more became aware. Unaware were not above chance for T (53%) or G (49%)
Rebuschat, Hamrick, Sachs, Riestenberg & Ziegler (2013)	Confidence rating after each test item: complete guess, somewhat confident, very confident, or absolute certainty. What basis selected: guess, intuition, memory, or rule knowledge Debriefing interview: what basis chose the determiners; is there a rule?; there is a rule, what could it be?; explain rule, did it occur?	9/15 expressed some awareness Experimental: 74.7% above chance (T: 78.3%; PT: 73.3%; N: 72.7% - all above chance) Control: above chance on new Sig group*item interaction 9 some awareness: sig above chance (total: 79.6%; T: 82.4%; PT: 79.6%; N: 76.9%), unaware not sig above chance (total: 53.5%; T: 64.6%; PT: 54.2%; N: 56.3%) Independent t-test: sig difference between aware and unaware only for PT For source attributes, guess (66.7%) and intuition (75.4%) were above chance

	ID results	Analysis
Williams, (2005)	Linguists sig better than non-linguists L1 gendered language not sig better but studying gendered language was. PTSM not sig	One-sample t-test correlation
Hama & Leow (2010)		One-sample t-test ANOVA 2x2 repeated measures Animacy (correct vs. incorrect) and distance (correct vs. incorrect) as within factor measures The dependent variables were types of responses—namely, correct animacy and correct distance (CACD), correct animacy and incorrect distance (CAID), incorrect animacy and correct distance (IACD), and incorrect animacy and incorrect distance (IAID) Paired samples t-test between CACD and IACD
Faretta-Stutenberg & Morgan-Short (2011)	No difference with gendered vs. non-gendered languages spoken Years of education correlate with generalised only for Aware group	One-sample t-tests ANOVA repeated measures Paired-samples t-test and correlations for ID measures
Rebuschat, Hamrick, Sachs, Riestenberg & Ziegler (2013)	Linguistics didn't affect results	One-sample t-tests ANOVA 2x3 (experimental, control) and (T, PT, N)

Appendix K. Study 2: Power Analysis

I used the pilot-based predictions (see Table 16) for the power analysis simulations. The pilot-based predictions were then run through a series of simulations (n=1000) using R (R Core Team, 2019) to determine an appropriate sample size to achieve a power of 80% for each aspect (memory and generalised) of the two regression models (intentional and incidental). The effects were simulated on the logit scale as changes in log odds which were then transformed into probabilities and simulated from a binomial distribution of those probabilities. Log odds were calculated for the intercept and then for each of the predictor variables (for each ISI group, and memory vs. generalised). See Table 16 for the log odds. For the intentional condition, the probability of answering an answer correctly in the 0-day ISI, memory items (.53) was used as the intercept (b0), corresponding to .12 in log odds. For each subsequent ISI group and for generalised items, an additional log odds produced by the interaction was calculated. For example, for 1-day ISI memory items in the intentional condition, the difference in log odds was .33 ((log odds of .61 – log odds of .53) = (.44-.12). For the 1-day ISI for the intentionalgeneralised items, it was .078 ((log odds of intentional-generalised 1-day ISI – (log odds of intentional-memory 0-day ISI) + (log odds of intentional–memory 1-day ISI – log odds of intentional-memory 0-day ISI)+(log odds of intentional-generalised 0-day ISI – log odds of intentional-memory 0-day ISI) = .41 – ((.12)+(.33)+(-.12)) = .078. The simulated data was then run through a multivariate logistic regression model (Bates et al., 2015) with fixed and random factors was built using R (R Core Team, 2019). Fixed effects included the delaygroup (0, 1, 2, 4, 7), itemtype (generalised and trained/memory) and learningtype (intentional and incidental). The random factors were at participant and item level. Using the car::Anova() package, the following model was built: glmer(accuracy ~ (1 + itemtype + learningtype | participant) + (1 + delaygroup| stimuli) + delaygroup + learningtype + itemtype + delaygroup:learningtype + delaygroup:itemtype).

For research questions 1 to 3 a sample size of 200 participants achieved a power of 92.4%, 80.6% and 35.9% respectively. Hence, only research question 3 and 4 were considered exploratory.

Power Analysis Script

```
# 17/03/2021 # Introduction #### # Variables # Accuracy: correct / incorrect response to a test item (outcome) # Learning type: intentional vs. incidental (within-participants) # Delay group: 0,1,2,4,7 days (between-participants) # Item type: old(memory) vs. new(generalised) (within-participants) # Stimuli: 96 in total: 24 intentional-old, # 24 intentional-new, # 24 incidental-new, # 24 incidental-old # Participants: balanced across delay group (must be factor of 5) # Research Questions # RQ1 # The optimal spacing group will be better than non-optimal, # irrespective of whether the condition is intentional or incidental # and whether the item-type is memory or generalised # RQ2 # intentional should have a wider optimal spacing than incidental, # irrespective of whether the item-type is memory or generalised # RQ3 # memory/old should have a wider optimal spacing than generalisation/new, # irrespective of whether it is in the intentional or incidental learning condition # Analysis # Each of the research questions will be addressed using an Anova() test of the # fixed effects in the model: # RQ1 - Is there a significant effect of delay group on accuracy, across learning # type and item type conditions # RQ2 - Is there a significant interaction between learning type and delay group, # across item type conditions # RQ3 - Is there a significant interaction between item type and delay group, # across learning type conditions # As each hypothesis test supports a single research question, alpha will be # fixed at 0.05 as the threshold for significance. Any post-hoc contrasts # will be adjusted for family-wise error rates by research question. Power # will be calculated as the probability of observing the significant effects # of interest for each research question, as defined above. # The power analysis will be based on simulations of the expected response data, # informed previous research and pilot data. Binary response accuracy will be # used at the outcome. The probability of achieving accuracy in each condition # is as follows: # 0.52 days0 incidental new # 0.50 days0 incidental old # 0.50 days0 intentional new # 0.53 days0 intentional old # 0.69 days1 incidental new # 0.58 days1 incidental old # 0.60 days1 intentional new # 0.61 days1 intentional old # 0.66 days2 incidental new # 0.69 days2 incidental old # 0.71 days2 intentional new # 0.73 days2 intentional old # 0.60 days4 incidental new # 0.61 days4 incidental old # 0.68 days4 intentional new # 0.77 days4 intentional old # 0.50 days7 incidental new # 0.50 days7 incidental old # 0.59 days7 intentional new # 0.67 days7 intentional old # Libraries #### library(dplyr) library(lme4) library(car) # Simuation Script #### # The following loop simulates a dataset according to the design and model (multilevel logistic) # with the selected parameter values, conducts the analysis and records the significance # of the 3 hypothesis tests described above. nsimulations = 1000 results = data.frame(matrix(nrow=0,ncol=0)) for(i in 1:nsimulations){ # Data frame construction nparticipant = 200 # total number of participants in across all groups (40 per delay group) nstimuli = 96 # total number of stimuli in all conditions (participants see all stimuli) ndelaygroups = 5 # Basic structure of data frame for participants and items simdata = data.frame(participant = rep(1:nparticipant, each = nstimuli), stimuli = rep(1:nstimuli, times = nparticipant)) # Add indicator for delay group simdata$delaygroup = rep(c("days0", "days1", "days2", "days4", "days7"), each = (nparticipant/ndelaygroups)*nstimuli) # Add indicator for learning type and item type simdata$learningtype = rep(c("incidental", "intentional"), each = nstimuli/2, times = nparticipant) simdata$itemtype = rep(c("new", "old"), each = nstimuli/4, times = nparticipant*2) # Check data frame construction: unique(simdata[3:5]) # Generating simulated data # Assuming the following
```

```
probabilities for each outcome: probabilitiesdf = data.frame(unique(simdata[3:5])) probabilitiesdf$probability = c(0.52, # days0 incidental new 0.5, # days0 incidental old 0.5, # days0 intentional new 0.53, # days0 intentional old 0.69, # days1 incidental new 0.58, # days1 incidental old 0.6, # days1 intentional new 0.61, # days1 intentional old 0.66, # days2 incidental new 0.69, # days2 incidental old 0.71, # days2 intentional new 0.73, # days2 intentional old 0.6, # days4 incidental new 0.61, # days4 incidental old 0.68, # days4 intentional new 0.77, # days4 intentional old 0.5, # days7 incidental new 0.5, # days7 incidental old 0.59, # days7 intentional new 0.67 # days7 intentional old ) # Convert probabilities to logits # Function to convert probability to log odds units: ptolo = function(p){ lo = -1*(log((1-p)/p)) # convert p into log odd units return(lo) } probabilitiesdf$logits = ptolo(probabilitiesdf$probability) # Combine probabilities and logits with simdata data frame simdata = left_join(simdata,probabilitiesdf,by=c("delaygroup", "learningtype", "itemtype")) # Simulate random intercept variances for participants simdata$pptintvar = rep(rnorm(nparticipant, 0, 1), each = nstimuli) # Simulate random intercept variances for items simdata$temintvar = rep(rnorm(nstimuli, 0, 1), times = nparticipant) # Simulate random slopes on within-participant effects # (item_type + learning_type + item_type:learning_type) # assuming simple contrast coding for simulation # reference levels for within-participant effects are: # learningtype: baseline = incidental # itemtype: baseline = new # Slope variance will be added for the contrast levels intentional and old # each participant will have a difference that will be repeated where # learningtype = intentional and itemtype = old, and 0 otherwise pptitemvar = rep(rnorm(nparticipant, 0, 1), each = nstimuli) simdata$pptitemvar = ifelse(simdata$learningtype == "intentional", pptitemvar, 0) # Check data frame: #unique(simdata[,10]) #unique(simdata[,c(1,4,10)]) pptlearningvar = rep(rnorm(nparticipant, 0, 1), each = nstimuli) simdata$pptlearningvar = ifelse(simdata$itemtype == "old", pptlearningvar, 0) # Check data frame: #unique(simdata[,11]) #unique(simdata[,c(1,5,11)]) # Combine defined logits with random variances simdata$outcomelogits = simdata$logits + simdata$pptintvar + simdata$temintvar + simdata$pptitemvar + simdata$pptlearningvar # Convert outcome logits to probabilities # Function to convert log odds to probability: lotop = function(lo){ p = 1/(1+exp(-lo)) # exp the log odds units return(p) } simdata$outcomeprob = lotop(simdata$outcomelogits) # Simulate observed accuracy (participant responses) simdata$accuracy = rbinom(1:nrow(simdata), size = 1, prob = simdata$outcomeprob) ## Plots #### # Visualising effects (omitting random variance) # library(ggplot2) # Probabilities #ggplot(simdata) + geom_point(aes(x=delaygroup, y=probability, colour=itemtype)) + # geom_line(aes(x=delaygroup, y=probability, colour=itemtype, group =itemtype)) + # facet_wrap(~learningtype) #ggplot(simdata) + geom_point(aes(x=delaygroup, y=probability, colour=learningtype)) + # geom_line(aes(x=delaygroup, y=probability, colour=learningtype, group =learningtype)) + # facet_wrap(~itemtype) # Logits: #ggplot(simdata) + geom_point(aes(x=delaygroup, y=logits, colour=itemtype)) + # geom_line(aes(x=delaygroup, y=logits, colour=itemtype, group =itemtype)) + # facet_wrap(~learningtype) #ggplot(simdata) + geom_point(aes(x=delaygroup, y=logits, colour=learningtype)) + # geom_line(aes(x=delaygroup, y=logits, colour=learningtype, group =learningtype)) + # facet_wrap(~itemtype) # Visualising the simulated random slopes and intercepts #sumlogits = data.frame(xtabs(outcomelogits ~ participant+learningtype+itemtype, simdata)) #sumlogits$avlogit = dummy$Freq/24 #sumlogits$delaygroup= rep(c("days0", # "days1", # "days2", # "days4", # "days7"), # each = (nparticipant/ndelaygroups)) # Item type: #ggplot(sumlogits, aes(group=participant)) + # geom_point(aes(y=avlogit, x = itemtype)) + # geom_line(aes(y=avlogit, x = itemtype, group = participant)) + # facet_wrap(~delaygroup*learningtype) # Learning type: #ggplot(sumlogits, aes(group=participant)) + # geom_point(aes(y=avlogit, x = learningtype)) + # geom_line(aes(y=avlogit, x = learningtype, group = participant)) + # facet_wrap(~delaygroup*itemtype) # Analysis Model #### model = glmer(accuracy ~ (1 + itemtype +
```

```
learningtype | participant) + (1 | stimuli) + delaygroup + learningtype + itemtype +
delaygroup:learningtype + delaygroup:itemtype, data = simdata, family = binomial(link="logit"))
htests = Anova(model, type = "III") p.values = data.frame("nsimulation" = i, "p.delaygroup" =
htests$`Pr(>Chisq)`[2], "p.learningXgroup" = htests$`Pr(>Chisq)`[5], "p.itemXgroup" =
htests$`Pr(>Chisq)`[6]) p.values results = rbind(results, p.values) print(paste("Progress: ",
i/nsimulations*100, "%", sep="")) } # Power Estimate #### # RQ1 rq1significance =
ifelse(results$p.delaygroup < 0.05, 1, 0) mean(rq1significance)*100 # RQ2 rq2significance =
ifelse(results$p.learningXgroup < 0.05, 1, 0) mean(rq2significance)*100 # RQ3 rq3significance =
ifelse(results$p.itemXgroup
```

Appendix L. Study 2: Exposure and Test Sentences

Item type	Intentional or Incidental	Determiner	Noun	Sentence
Training - old		gi	bear	I tried to play dead while gi bear sniffed me.
Training - old		ul	bee	The child screamed even though ul bee was on the other side of the playground.
Training - old		ro	box	I carefully packed my nephew's present in ro box with bubble wrap.
Training - old		ne	book	I couldn't read the title of ne book that was on the top shelf.
Training - old		ne	plate	From the kitchen I heard the sound of ne plate crashing to the floor.
Training - old		gi	dog	To reward him for his trick, the owner fed gi dog a treat.
Training - old		gi	pig	I held gi pig tightly so he wouldn't escape from the pen.
Training - old		ul	cat	The old lady pointed at ul cat stuck at the top of the tree.
Training - old		ul	fly	I could see ul fly on the projector screen from the back of the classroom.
Training - old		ro	picture	Looking closely, the detective checked ro picture for clues about the crime.
Training - old		ro	sofa	The movers carefully carried ro sofa up the stairs.

Training - old	ne	cushion	I stood back to see how well ne cushion matched the couch.
Training - old	gi	rat	While I was sitting in the kitchen gi rat ran between my feet.
Training - old	ul	bird	The naturalist tried to take a picture of ul bird feeding on the other side of the lake.
Training - old	ro	cup	I knocked over ro cup and the coffee spilled on my book.
Training - old	gi	lion	I was terrified when I turned around and saw gi lion right behind me.
Training - old	gi	cow	The vet carefully examined gi cow at the farm.
Training - old	ul	monkey	The children threw sticks at ul monkey in the tree.
Training - old	ul	snake	We could hear ul snake hissing from behind the bushes.
Training - old	ro	table	I spent an hour polishing ro table before the dinner party.
Training - old	ro	television	The girl had to switch off the ro television manually as the remote was missing.
Training - old	ne	clock	I looked up at ne clock on the church and realized that I was late.
Training - old	ne	stool	In the pub I asked my friend to get ne stool from the bar.
Training - old	ne	vase	I wanted to see ne vase from China at the far end of the museum.
Training - new	gi	lizard	He liked the feel of the scales of gi lizard.
Training - new	ul	mouse	The kitchen porter spotted ul mouse run behind the oven.
Training - new	ro	bottle	I picked up ro bottle and poured two glasses of wine.

Training - new	ne	umbrella	I had left ne umbrella at home so I got wet.
Training - new	gi	eagle	The vole was snatched up by gi eagle that had swooped in for the kill.
Training - new	ul	penguin	From across the water, the explorer watched ul penguin slide along the ice shelf.
Training - new	ro	coin	I asked my nephew to guess which hand ro coin was in.
Training - new	ne	key	The man asked his son to fetch ne key from the front door.
Training - new	gi	donkey	The child went for a ride on gi donkey on the beach.
Training - new	ul	rhino	From so far away, the tourist couldn't tell how big ul rhino was.
Training - new	ro	postcard	I sat in a café and wrote ro postcard to my family.
Training - new	ne	door	The police officers sat in their car and watched ne door, waiting for the gang to come out.
Training - new	gi	gorilla	The park ranger sat next to gi gorilla for several hours.
Training - new	ul	chimpanzee	I watched as ul chimpanzee swung from the branch of the tree.
Training - new	ro	toothbrush	The girl carefully added some toothpaste to ro toothbrush.
Training - new	ne	mirror	He stood back and looked in ne mirror to check out his new clothes.
Training - new	gi	crocodile	Captain Hook had his hand bitten off by gi crocodile.
Training - new	ul	cheetah	Watching ul cheetah accelerate from 0 to 60 miles per hour was incredible.
Training - new	ro	battery	She took ro battery out of the remote and threw it away.

Training - new		ne	newspaper	The road sweeper saw that ne newspaper was all over the street.
Training - new		gi	hedgehog	As soon as she picked up gi hedgehog, it curled into a ball.
Training - new		ul	racoon	I could hear ul racoon running around in the attic.
Training - new		ro	comb	The rocker kept ro comb in his shirt pocket.
Training - new		ne	bowl	In the middle of the large table was ne bowl that we'd given them as a present.
	Int or Inc			
TEST - trained	Int	ro	cup	The babysitter poured juice into __ cup for the child.
TEST - trained	Inc	ul	cat	I heard the sound of ul cat meowing in the old barn.
TEST - trained	Int	ul	monkey	I turned around and saw __ monkey scampering away with my banana.
TEST - trained	Inc	ro	television	I moved ro television aside to make room for the new speakers.
TEST - trained	Int	gi	bear	After years of abuse, __ bear finally attacked the cruel circus ringmaster.
TEST - trained	Inc	ne	vase	On the other side of the room the wind blew the window open knocking __ vase off the sill.
TEST - trained	Int	ne	book	I ordered ne book from the library online.
TEST - trained	Inc	ne	cushion	When I had a backache I asked my wife to fetch __ cushion from the bedroom.
TEST - trained	Int	ro	table	The child asked if he could get down from __ table as he had finished eating.
TEST - trained	Inc	ro	box	The girl tore open __ box to get at the cookies inside.

TEST - trained	Int	gi	cow	The farmer was kicked by __ cow when he tried to milk it.
TEST - trained	Inc	ne	clock	The tourists had trouble reading __ clock on the tower across the river.
TEST - trained	Int	ul	snake	He was alarmed to see __ snake on the other side of the garden.
TEST - trained	Inc	gi	lion	In the widlife sanctuary you can pay extra to stroke __ lion cub.
TEST - trained	Int	ul	bird	The guide set up the telescope so we could see __ bird.
TEST - trained	Int	ne	plate	At the fair, the child threw balls at __ plate to win a prize.
TEST - trained	Inc	gi	pig	The flies swarmed around the head of __ pig.
TEST - trained	Inc	gi	rat	The chef tried to hit __ rat with his rolling pin.
TEST - trained	Int	ne	stool	We hoped that __ stool would be delivered in time for the party.
TEST - trained	Inc	ul	bee	The researcher studied __ bee from a safe distance.
TEST - trained	Inc	gi	dog	When I was out for a walk I stopped to stroke __ dog and it bit me.
TEST - trained	Int	ro	picture	The art museum curator handled __ picture for signs of damage.
TEST - trained	Inc	ro	sofa	I spent the night on ro sofa and let my guests sleep in the beds.
TEST - trained	Inc	ul	fly	The man had to squint to see ___ fly over on the wall.
TEST - trained	Int	gi	lizard	He jumped when __ lizard ran over his foot.

TEST - trained	Int	ul	mouse	The girl heard __ mouse scurrying in the floorboards above her head.
TEST - trained	Int	ro	bottle	In the wine cellar, the millionaire handed her guest __ bottle that cost £10,000.
TEST - trained	Inc	ne	umbrella	He was determined to buy __ umbrella that he'd seen in the shop earlier that day.
TEST - trained	Int	gi	eagle	In the bird of prey sanctuary, he had __ eagle to perch on his arm.
TEST - trained	Inc	ul	penguin	The leopard seal saw __ penguin swimming a long way away.
TEST - trained	Inc	ro	coin	When I was on holiday I couldn't tell the denomination of __ coin in my hand.
TEST - trained	Int	ne	key	The traveller could see __ key hanging up behind the receptionist on the front desk of the hotel.
TEST - trained	Int	gi	donkey	My wife gave __ donkey an apple.
TEST - trained	Inc	ul	rhino	They sat in the jeep and waited as __ rhino crossed the track 100 metres ahead of them.
TEST - trained	Int	ro	postcard	I picked out ___ postcard from the rack and took it to the cashier.
TEST - trained	Inc	ne	door	The man was sleeping in bed when __ door downstairs burst open.
TEST - trained	Inc	gi	gorilla	When the child fell into the enclosure, __ gorilla picked him up and carried him to the entrance.
TEST - trained	Int	ul	chimpanzee	The girl waved at __ chimpanzee in the enclosure and was amazed when it waved back.
TEST - trained	Inc	ro	toothbrush	The battery in __ toothbrush ran out half way through brushing his teeth.
TEST - trained	Int	ne	mirror	Looking up, he was amazed to see ___mirror on the high ceilings.

TEST - trained	Inc	gi	crocodile	The tiny bird landed on __ crocodile thinking it was a log.
TEST - trained	Int	ul	cheetah	Looking through my binoculars I could see __cheetah lying under a tree.
TEST - trained	Inc	ro	battery	She had to use a pair of scissors to get __ battery out of its packet.
TEST - trained	Int	ne	newspaper	I couldn't make out the headline of __ newspaper at the stand across the street.
TEST - trained	Int	gi	hedgehog	When he lifted the lid on the feeding station, __ hedgehog was right there.
TEST - trained	Inc	ul	racoon	When he heard the dustbin outside crash over, he knew it was __racoon forraging for scraps.
TEST - trained	Int	ro	comb	The barber used __comb to create a fine parting in the customer's hair and then he added gel.
TEST - trained	Inc	ne	bowl	The mother told her daughter to get __ bowl from the kitchen if she wanted ice-cream.
TEST - new	Int	gi	hamster	He picked __ hamster up and put it in its wheel.
TEST - new	Inc	ul	turtle	From the path the hikers saw ul turtle sunning himself on a rock.
TEST - new	Int	ul	camel	The photographer took a stunning photo of __ camel across the sand dunes.
TEST - new	Inc	ro	desk	The teacher made the naughty student sit at __ desk right in front of her.
TEST - new	Int	ro	phone	I was surprised when ro phone rang in my hand.
TEST - new	Inc	ne	candle	From the window of his room, the monk could see __ candle moving quickly across the courtyard
TEST - new	Int	ne	lamp	She asked her husband to turn off __ lamp in the other room.

TEST - new	Inc	ul	horse	I couldn’t see __ horse in the field without glasses.
TEST - new	Inc	gi	rabbit	The child held gi rabbit at the petting zoo.
TEST - new	Int	ro	spoon	The man put away ___ spoon in the drawer.
TEST - new	Inc	gi	elephant	The kids at the zoo fed __ elephant peanuts right from their hands.
TEST - new	Int	ne	towel	The heiress asked the pool boy to get __ towel from the rack.
TEST - new	Int	gi	otter	The zookeeper fed __ otter the fish by hand.
TEST - new	Int	ul	squirrel	I couldn't take a picture of __ squirrel as it kept running behind the tree.
TEST - new	Inc	ro	rug	It felt so comfortable when my toes sank into __ rug.
TEST - new	Inc	ne	bench	The headmaster asked the boy to leave the office and wait on __ bench outside.
TEST - new	Int	gi	swan	The child had his arm broken by __ swan.
TEST - new	Inc	ul	beaver	The mother told her son if he looked carefully he would be able to see __ beaver building its dam.
TEST - new	Int	ro	bin	I hate emptying __ bin. My hands get dirty.
TEST - new	Int	ne	shoe	The little boy pointed up at __ shoe that he wanted on the shop wall.
TEST - new	Inc	gi	ferret	They gave the elderly patient ___ ferret to stroke as a form of therapy.
TEST - new	Int	ul	lobster	She could see ___ lobster in the lobster cage as the fisherman raised it from the water.

TEST - new	Inc	ro	brick	The builder put the mortar around the edge and then pushed ___ brick gently in the gap.
TEST - new	Inc	ne	knife	As soon as the detective entered the room, he saw ___ knife lying on the floor covered in blood.
TEST - new	Inc	gi	snail	The gardener picked __ snail off his lettuce.
TEST - new	Int	ul	deer	The hunter caught a glimpse of __ deer before it disappeared into the forest.
TEST - new	Int	ro	flag	The Amierican soldiers ceremoniously folded up __ flag and buried it.
TEST - new	Inc	ne	ball	My brother told me to go and get __ ball after I had kicked it into our neighbour's garden.
TEST - new	Int	gi	bat	It was really scary when __ bat flew into my hair.
TEST - new	Inc	ul	octopus	From the boat we could see __ octopus swimming beneath the surface.
TEST - new	Inc	ro	bathtub	She filled __ bathtub with water and added some bubble bath.
TEST - new	Int	ne	shovel	The boy remembered __ shovel was still outside.
TEST - new	Int	gi	shark	The diver was scared stiff when __ shark swam right by her.
TEST - new	Inc	ul	dolphin	From the beach we could see __ dolphin jump from the water and splash its tail.
TEST - new	Int	ro	wok	The chef added some garlic to __ wok and stir-fried it for thirty seconds.
TEST - new	Inc	ne	paintbrush	The young artist longed for __ paintbrush she saw through the window of the art supply shop.
TEST - new	Inc	gi	llama	The farmer sheared __ llama of its wool

TEST - new	Int	ul	kangaroo	He saw __ kangaroo bouncing off down the road.
TEST - new	Inc	ro	basket	Little Red Riding Hood put flowers in __ basket.
TEST - new	Int	ne	ring	She tried desperately to remember where she'd left __ ring.
TEST - new	Int	gi	spider	While the camper was sleeping __ spider crawled across his face.
TEST - new	Inc	ul	badger	The headlights of the car briefly lit up __ badger as it scurried along.
TEST - new	Int	ro	spatula	Using __ spatula he carefully flipped the eggs.
TEST - new	Inc	ne	corkscrew	At the picnic he couldn't open the wine because __ corkscrew was at home.
TEST - new	Inc	gi	wasp	It hurt like crazy when he was stung by __ wasp.
TEST - new	Int	ul	owl	He heard the hooting of __ owl in the woods.
TEST - new	Inc	ro	pillow	As soon as the girl lay her head on __ pillow, she fell asleep.
TEST - new	Int	ne	bag	My wife asked me to go out and get __ bag from the car.

Note. Exposure = items used in the exposure phase; Test= items used in the delayed posttest; old = items taken from Rebuschat et al. (2013);

new = items created for the current study; Intentional/Incidental = aspect of the determiner that is tested in the test item.

Appendix M. Study 2: Instructions and Sample Exposure Questions (Animacy as Intentional Aspect)

Q1
You are going to learn four new words
ro, gi, ul, ne

ro = non-living ne = non-living
ul = living gi = living

Study their meanings for thirty seconds. On the next page you will be tested.

- ○ Click here when you are ready to be tested (281)

JS

Q2

ro = ?

- ○ living (1)
- ○ non-living (4)

JS

Q3

ul = ?

- living (1)
- non-living (4)

Carry Forward All Choices - Displayed & Hidden from " ul = ?"

JS X→

Q4

ne = ?

- living (1)
- non-living (2)

Carry Forward All Choices - Displayed & Hidden from " ne = ?"

JS X→

Q5

gi = ?

- living (1)
- non-living (2)

End of Block: Pre-teach vocab ne, ro, ul, gi Intentional Animacy

---

Start of Block: Pre-teach vocab repeat you got one wrong

JS

Q1
You got at least one wrong. Let's try again. Spend 30 seconds on the next page trying to remember the meanings of the four words.

Click here when you are ready (1)

End of Block: Pre-teach vocab repeat you got one wrong

Start of Block: Pre-teach vocab end - 3 strikes and you're out

Q1
Unfortunately, you did not get them right. You are therefore not suitable for this experiment. Thank you very much for your interest in this study.

Click when you are ready (7)

End of Block: Pre-teach vocab end - 3 strikes and you're out

Start of Block: Training instructions Intentional Animacy

JS

Q1 This is an experiment about how people formulate sentences in different languages. Different languages have different ways of saying things, which raises the question of whether this forces people to think slightly differently in order to speak and understand in different languages.
Here I am investigating this issue in a situation where the sentences are almost entirely in English, apart from the four words that you have just learned.

Click here when you've finished reading (1)

JS

Q2
In this language, each time an object is mentioned, it is necessary to specify whether it is "living" or "non-living". The word that is used to do this also functions like the English word “the”. So, saying "gi tiger" is like saying “the-living tiger”.

Click here when you've finished reading (1)

Carry Forward Selected Choices from "In this language, each time an object is mentioned, it is necessary to specify whether it is "living" or "non-living". The word that is used to do this also functions like the English word "the". So, saying "gi tiger" is like saying "the-living tiger"."

JS X→

Q3 In the context of the example

"The girl is patting gi tiger"

the word "gi" expresses the idea that the tiger is living -obviously, because it is an animal!

- Click here when you've finished reading (1)

JS

Q4
Let's practice. Read the whole sentence silently. Make a mental picture of the sentence. Click on whether the noun after *ne* is living or non-living

The man and girl saw ne church in the distance.

- living (1)
- non-living (4)

JS

Q5
The man and the girl saw ne church in the distance.

In this sentence the word "ne" expresses the idea that the church is non-living -obviously, because it is a building!

- Click here when you've finished reading (1)

JS

Q6
Okay, we are about to start. In the following section, you should:
a) read the whole sentence silently
b) make a mental picture of the sentence
c) click whether the noun affer *ul/ne/gi/ro* is living or non-living
Please ensure you do all three.

- Click here to start (1)

End of Block: Training instructions Intentional Animacy

Start of Block: Training 48 sentences -Intentional - Animacy block 1

JS

Q1 I tried to play dead while gi bear sniffed me.

- living (1)
- non-living (2)

JS

Q2
The child screamed even though ul bee was on the other side of the playground.

- living (1)
- non-living (2)

JS

Q3
I carefully packed my nephew's present in ro box with bubble wrap.

- ○ living (1)
- ○ non-living (2)

Page Break

Appendix N. Study 2: Testing Block Instructions and Sample Test Questions (Animacy as Intentional Aspect)

JS

Q18
Test phase
In this part you need to choose the best word to go in the gap.

- Click here to start

End of Block: Test setup

Start of Block: Test (96)

JS

Q9 The babysitter poured juice into __ cup for the child.

- gi
- ro

JS

Q95 I heard the sound of __ cat meowing in the old barn.

- ul
- ne

JS

Q96 I turned around and saw __ monkey scampering away with my banana.

- ul
- ne

JS

Q97 I moved __television aside to make room for the new speakers.

- ro
- ne

JS

Q98 After years of abuse, __ bear finally attacked the cruel circus ringmaster.

- gi
- ro

JS

Q99 On the other side of the room __ vase fell off the window sill.

- ro
- ne

Appendix O. Study 2: Debriefing Questionnaire

Q193 Thank you. That is the end. In this experiment you had to choose from four words (gi, ro, ne and ul). Did you notice any rules for their use (including any that you were told about at the beginning of the experiment)? For each rule you noticed, fill in a different text box. Please make guesses if you aren't sure. You don't need to fill them all in.

- ☐ Rule 1 (1) _____
- ☐ Rule 2 (12) _____
- ☐ Rule 3 (13) _____
- ☐ Rule 4 (14) _____
- ☐ Rule 5 (15) _____
- ☐ Rule 6 (16) _____
- ☐ Rule 7 (17) _____
- ☐ Rule 8 (18) _____
- ☐ I didn't notice any rules (19)

JS

194 When did you notice the rule that ${Q193/ChoiceTextEntryValue/1}

- When the instructions were given (1)
- During day 1 practice (2)
- Between day 1 practice and day 2 practice (3)
- During day 2 practice (4)
- Between day 2 practice and day 3 practice (5)
- During day 3 practice (6)
- Between day 3 practice and the delayed test (7)
- During the delayed test (8)
- When asked about rules just now (9)

JS

Q198 How sure are you of the rule that ${Q193/ChoiceTextEntryValue/1} ?

- Not sure at all (1)
- Not sure (2)
- I think so (3)
- I'm very sure (4)

JS X→

Q195 When did you notice the rule that ${Q193/ChoiceTextEntryValue/12}

- When the instructions were given (1)
- During day 1 practice (2)
- Between day 1 practice and day 2 practice (3)
- During day 2 practice (4)
- Between day 2 practice and day 3 practice (5)
- During day 3 practice (6)
- Between day 3 practice and the delayed test (7)
- During the delayed test (8)
- When asked about rules just now (9)

Q199 How sure are you of the rule that ${Q193/ChoiceTextEntryValue/12} ?

- Not sure at all (1)
- Not sure (2)
- I think so (3)
- I'm very sure (4)

Page Break

Q196 When did you notice the rule that ${Q193/ChoiceTextEntryValue/13} ?

- When the instructions were given (1)
- During day 1 practice (2)
- Between day 1 practice and day 2 practice (3)
- During day 2 practice (4)
- Between day 2 practice and day 3 practice (5)
- During day 3 practice (6)
- Between day 3 practice and the delayed test (7)
- During the delayed test (8)
- When asked about rules just now (9)

Q200 How sure are you of the rule that ${Q193/ChoiceTextEntryValue/13} ?

- Not sure at all (1)
- Not sure (2)
- I think so (3)
- I'm very sure (4)

Page Break

Q202 1. If you did not notice a rule for when to choose between gi and ul, or when to choose between ro and ne, make a guess at a rule now.

- ○ Rule: gi and ro are both... (1) _____
- ○ Rule: ul and ne are both... (2) _____
- ○ I still can't find a rule (3)

JS

Q203 One pair of words refers to near things and the other pair refer to far things.
Which one does *gi* and *ro* refer to?

- ○ Near (1)
- ○ Far (2)

JS

Q204 Did you write down anything regarding the rules in between study sessions?

- ○ Yes (1)
- ○ No (2)

Q211 In the exposure phases of this experiment, did you read the sentences aloud?

- ○ Yes (4)
- ○ No (5)

Q205 What do you think the purpose of this experiment was?

○ . (1) _____

Q212 Do you have any comments about this experiment and how it was run?

_____

End of Block: Debriefing questionnaire

Start of Block: MLAT post-test

JS