# Does spacing retrieval practice lead to a benefit in word learning before and after sleep? A developmental perspective

Maja Amanda Olsson

Doctor of Philosophy

University of York
Psychology

September 2022

# Abstract

Understanding how to promote word learning across development is a crucial research area, as vocabulary size can predict academic and employment success. Spaced (relative to massed) retrieval practice (i.e., practising retrieval in sessions separated by hours or days) has been argued to benefit the long-term retention of novel vocabulary in adults and children. However, the effects of spaced retrieval practice are similar to sleep effects and the effects of repeated retrieval practice with feedback. Few studies have examined spaced retrieval practice in the absence of sleep between retrieval attempts or without feedback. This thesis addresses these areas by examining word learning performance up to 7 days after within-day spaced or massed retrieval practice with or without feedback and with or without a retrieval opportunity 24 hours after initial training in adult and child populations. We found a benefit in word learning performance from massed retrieval practice in adults and children on the first day of testing. Further, while within-day spaced retrieval practice protected memories from forgetting one week later, novel words were not better retained after within-day spaced retrieval practice in adult or child populations. We also conducted a direct comparative analysis of the adult and child data where we found that within-day spaced retrieval practice or a post-sleep retrieval opportunity generally slowed forgetting in adults, but children continued to improve their performance. However, developmental differences in the effects from within-day spaced retrieval practice only emerged if a post-sleep retrieval opportunity was provided. These results will be considered in the context of retrieval effort, contextual variability, and reconsolidation theories. Overall, this thesis provides critical evidence that a spacing benefit does not always emerge and that a post-sleep retrieval opportunity slows the forgetting of novel words one week later in adults but encourages continued improvements in children.

# Table of Contents

## List of Tables

## List of Figures

# List of Appendices

# Acknowledgements

First, my biggest and most grateful thank you goes to my fantastic supervisors, Dr Lisa Henderson and Prof Gareth Gaskell. I cannot thank you enough for all your support, patience, and encouraging words that kept me going whenever I lost my confidence. I am incredibly grateful for the countless hours you have dedicated to giving me feedback and for always pushing me to do my best. It is easy to say that without your constant support, I would not have finished my PhD with five experiments that I am very proud of.

I have been lucky to have Dr Tom Hartley and Prof Asifa Majid as part of my thesis advisory panel. Thank you for listening to my presentations, engaging in insightful discussions, and providing general support throughout my PhD journey.

I also want to thank all the Sleep Language and Memory lab members for giving me a space to present my ideas and ask questions. Specifically, I want to thank the members of the SleepSmart project: Dr Lisa Henderson, Dr Fay Fletcher, Dr Sarah Walker, Dr Victoria Knowland, and Dr Elaine van Rijn. Thank you for taking me on as a research assistant and involving me with the data collection of many of your projects. Your passionate approach to research sparked my own love for research and data collection.

To my participants, thank you for dedicating your time to help us better understand the spacing effect. Special thanks to the participating schools for letting me run my studies during the very turbulent time of the Covid-19 pandemic.

Thank you to my friends Matt Foxwell, Nicholas Souter, Emma Raat, Emma Jackson, and Federico Segala for inviting me to play Dungeons and Dragons with you through our PhD journeys. Hunting dragons and starting goblin rebellions with you have been the perfect way to relax even during the most stressful periods of my PhD. A special thank you goes to Matt

Foxwell for spending so much of your time preparing our sessions. And thank you for introducing me to Mell and Zell. They will continue to explore forests and save animals long after we have finished our PhDs.

My final thanks go to my family and my partner. Thank you to my parents, Sara and Håkan, and my brother, Adam, for always being there and for patiently listening to me trying to explain my research in Swedish. I am proud to have such an awesome family! *Tack för allt ert stöd, ni är bäst! Och jag vet, jag är bäst!* To my wonderful partner, John, thank you for coming along on this PhD journey with me. I could not have done it without your constant support! Thank you for giving up plans so I could meet deadlines, for letting me practice presentations with you, and for always making me laugh through good and bad times. Having you by my side has made my PhD and life an adventure that I look forward to continuing together.

## Authors declaration

This thesis is a presentation of original work completed solely by the author under the supervision of Dr Lisa Henderson and Professor Gareth Gaskell. This work has not been previously presented for an award at this or any other university. All sources are acknowledged as references. The research was supported by a studentship at the Department of Psychology, University of York, UK.

### Conference presentations

Chapter 2:

Olsson, A., Gaskell, M.G., Henderson, L.M. (2020, June). *Feedback in Spaced or Massed Retrieval Practice Affects Word Learning in Adults.* Poster presentation at the International Max Planck Research School (IMPRS) conference for Language Sciences, online.

Olsson, A., Gaskell, M.G., Henderson, L.M. (2019, June). *The effectiveness of within-day spaced retrieval practice.* Poster presentation at the PGR conference at the University of York.

Chapter 4:

Olsson, A., Gaskell, M.G., Henderson, L.M. (2021, June). *Does spacing retrieval practice over a day lead to benefits in word learning in adults and children?* Presentation at a department-wide seminar, online.

# Chapter 1:

## General introduction - Can long-term retention of words be enhanced from spaced retrieval practice? Developmental differences and the role of feedback and sleep

## 1.1 Introduction

The ability to learn and remember is crucial across the lifespan, and finding ways to optimise learning and retention of new memories has been of interest to psychologists for decades (e.g. Spitzer, 1939). One critical process supported by our memory systems is word learning, which culminates in vocabulary knowledge. It has been found that the breadth and depth of an individual's vocabulary knowledge is a strong predictor of academic success in children (e.g. Biemiller, 2003; Fernald, Marchman, & Weisleder, 2013) and adults (e.g. Treffers-Daller & Milton, 2013), employment outcomes (e.g. Fernald, Marchman, & Weisleder, 2013; Law, Rush, Schoon & Parson, 2009), and even mental health (Armstrong et al., 2017). Thus, research that finesses methods of encouraging effective and long-lasting vocabulary throughout development has clear societal value, beyond having an essential role in advancing our theoretical understanding of memory systems in action.

There is a consensus that learning a new word is not quick or immediate but takes place over a substantial period. For example, there has been considerable interest in sleep's role in embedding a new word in the existing mental lexicon. Several methods have been suggested to help enhance different aspects of the word learning process, from initial encoding to long-term recall. Regarding the latter, two methods that have attracted long-standing interest in both adults and children are retrieval practice and spaced learning. Retrieval practice is when new information is actively recalled through tests such as cued recall or multiple-choice questions and contrasts with restudying, which usually involves passive exposure to the new information. Spaced learning is when learning sessions are distributed over extended periods, with gaps between sessions ranging from minutes to several days. Spaced learning is often contrasted with massed learning when learning occasions take place over a shorter time, often a single session (for example, when a student tries to learn a whole semester's worth of information the evening before an exam). Both retrieval practice and spaced learning have been argued to promote learning and long-term retention of learned words, especially when the two methods are combined (e.g. Carpenter, Cepeda, Rohrer, Kang & Pashler, 2012), henceforth referred to as *spaced retrieval practice.*

However, *how* the learning and retention of new words are influenced by spaced retrieval practice, including how this process is influenced by sleep and whether it differs between childhood and adulthood, remains unclear. Recently, an account that sleep may contribute to the long-term benefits of spaced retrieval practice has been put forward. The present review will examine this account's evidence and traditionally popular theoretical accounts for the spacing effect (i.e., retrieval effort and contextual variability theories). Further, given previous findings that the magnitude of sleep-based consolidation effects can differ through development, we also explore whether the role of spaced retrieval practice differs over development.

First, the review will consider the process of word learning in adults from the perspective that word learning draws on general cognitive memory systems influenced by sleep. With this theoretical backdrop, evidence for whether spaced retrieval practice can bolster adult word learning will be considered, and the conditions that may influence these effects. Then, the effectiveness of spaced retrieval practice in promoting long-term retention in children will be reviewed, along with considering the developmental stability of these effects and whether there are differences between adults and children.

## 1.2 Word learning and spaced retrieval practice in adults

### 1.2.1 Word learning in adults

Theoretical accounts of word learning have conceptualised the process as domain-specific (i.e. specific processes for word learning; e.g. Fodor, 1983) or domain-general (i.e. drawing on general cognitive memory systems; e.g. Palma & Titone, 2020). By approaching word learning as a domain-general process, researchers studying word learning can draw conclusions relating to word learning and the human memory system in general (Palma & Titone, 2021). This approach also allows a more nuanced discussion around word learning as a long and complex procedure that draws on several memory processes (e.g. Leach & Samuel, 2003). With this approach as a backdrop, this review will outline general cognitive memory systems, which have also been explicitly applied to word learning.

The ultimate goal of the word learning process is *consolidation*. That is when new memories go from a fragile state (at initial learning or *encoding*) to a stable state (allowing long-term recollection). Advocates of the domain-general complementary learning systems (CLS) account (McClelland et al., 1995; O'Reilly & Rudy, 2000; O'Reilly et al., 2014) propose that new memories are initially encoded in both hippocampal and neocortical brain structures. However, initially, memories rely more on the hippocampal structures. Over time, reliance transfers to long-term neocortical areas where memories are more accessible for later retrieval in a range of contexts.

Applying this CLS model to word learning, Davis and Gaskell (2009) theorised that novel words are initially encoded as events (i.e. episodic memories) in hippocampal and medial-temporal lobe structures. Memories in these regions are thought to be individually coded through rapid but sparse activation patterns (e.g. Yassa & Stark, 2011). Thus, the initial trace of a novel word might be easily associated with its learning conditions but not easily generalised, making it ineffective for everyday use. Over time, these novel words become integrated as distributed and overlapping representations within semantic memory in the neocortex, strengthening its memory trace and allowing the generalisation of the word and its meaning to a broad range of contexts. The overlapping semantic representations of words in the neocortex (e.g. superior temporal gyrus) are posed to form a "*mental lexicon*", allowing humans to use words efficiently and flexibly in language.

How do memories of new words go from being fragile, episodic representations in hippocampal structures to relying on neocortical structures as stable, semantic representations? According to CLS, this occurs during a consolidation process. Through consolidation, the hippocampal areas are thought to aid the organisation of new neocortical traces, allowing novel words to interleave with existing words in the mental lexicon. For this to happen, hippocampal areas must communicate with the neocortex through the reactivation of memory traces in a coordinated effort. The reactivation of relevant structures can occur online through activities such as active retrieval (e.g. Antony et al., 2017) or offline through sleep (McClelland et al., 1995; Davis & Gaskell, 2009).

While this review will focus on the role of sleep and retrieval practice in the consolidation process of novel words, there is an argument that novel information (including words) can be more rapidly integrated with neocortical areas if they are related to existing schemas (e.g. Tse et al., 2007, 2011; McClelland, McNaughton & Lampinen, 2020; McClelland, 2013). Thus, if a novel word is related to existing words, a period of consolidation may not always be required for integration in neocortical regions. However, which aspects need to be related for integration without sleep in word learning remains unclear. For example, a word can contain phonemic, semantic and orthographic information, and it is not clear if these aspects are equally important to be related for integration to occur without sleep.

#### *1.2.1.1 Sleep-based processes that aid consolidation of novel words*

Before understanding the role of sleep in word learning, it is essential to understand the basics of human sleep architecture. During a typical night of sleep, adult humans progress through different sleep stages in 90-minute cycles. The sleep stages are often separated into rapid eye movement (*REM*) and non-REM (*NREM*) sleep, with the first half of the night generally consisting of more NREM sleep and the second half consisting of more REM sleep (Rasch & Born, 2013). REM sleep is a very light form of sleep, characterised by fast and low-amplitude oscillatory brain activity similar to brain activity when waking. NREM sleep is often divided further into three distinct stages of sleep. Sleep stage 1 is the first stage of sleep between wakefulness and deeper stages of sleep. Stage 2 is deeper sleep than stage 1 and is characterised by slow oscillations (0.75Hz), K-complexes (high amplitude, sharp wave at 80-100Hz), and sleep spindles (brief bursts of 10-16Hz activity, lasting between 0.5-1.5 seconds; Staresina et al., 2015; Berry & Wagner, 2015). Finally, the deepest stage of sleep is stage 3, also called slow wave sleep (*SWS*), which is characterised by an abundance of high-amplitude *slow oscillations* (waves of 0.75Hz activity). The consolidation of new words has been associated with specific EEG activity that characterises each sleep stage, most commonly slow oscillations and spindles (e.g. Tamminen, Payne, Stickgold, Wamsley & Gaskell, 2010), but also including subsequent cycles of REM sleep (e.g. Tamminen, Lambon-Ralph & Lewis, 2017). This review will focus more on the neurological aspects of active systems consolidation that occur during NREM sleep.

SWS is particularly interesting when considering word learning as much activity during this stage (e.g. slow oscillations and spindles) has been linked to the consolidation of novel words (e.g. Born, 2010; Tamminen et al., 2010). As mentioned, this stage is characterised b*y* slow oscillations, which originate from neocortical areas (mainly prefrontal regions; Timofeev et al., 2000) and permeate across the entire neocortex, hippocampus and thalamic regions (Nir et al., 2011; Steriade, 2006). In addition, short bursts of spindle activity can occur temporally synchronised with slow oscillations (Staresina et al., 2015). There are two types of spindles, slow (<12Hz) and fast (>12Hz), with fast spindles occurring in the downstate of slow oscillations and are thought to encourage communication between neocortical and hippocampal regions (Staresina et al., 2015). Spindles are triggered in thalamic regions and spread into the neocortex (Steriade, 2006) and hippocampal areas (e.g. Sarasso et al., 2014). Relevant to the consolidation process, where communication between the hippocampus and neocortex is critical, hippocampal spindles continuously precede spindles in neocortical regions (Sarasso et al., 2014), suggesting temporally coordinated activation. Also relevant to this process are *hippocampal ripples* (80-100Hz), which are temporally synchronised with the downstate of spindles and are thought to signal the reactivation of memory traces (Staresina et al., 2015; Wilson & McNaughton, 1994).

Thus, in line with the CLS model, during NREM in humans, slow oscillations occur, triggering sleep spindles in hippocampal and neocortical areas. The coordinated spindle activation in hippocampal and neocortical areas has been suggested to indicate that the areas communicate information with each other (Ngo, Fell & Staresina, 2020; Demanuele et al., 2017; Mölle  & Born, 2011; Mölle, Marshall, Gais & Born 2002). Specifically, hippocampal ripples represent reactivations of memories in the hippocampus, which are then communicated to neocortical areas integrating the words into the mental lexicon allowing long-term retention (Ngo, Fell & Staresina, 2020; Staresina et al., 2015). Furthermore, there is evidence for a causal relationship between slow oscillations and coordinated memory processes, as boosting the activity of slow oscillations using transcranial direct current stimulation during sleep can improve the later retention of memories (Marshall, Helgadottir, Mölle & Born 2006). Figure 1 illustrates how these processes encourage communication across cortical regions to promote long-term retention.

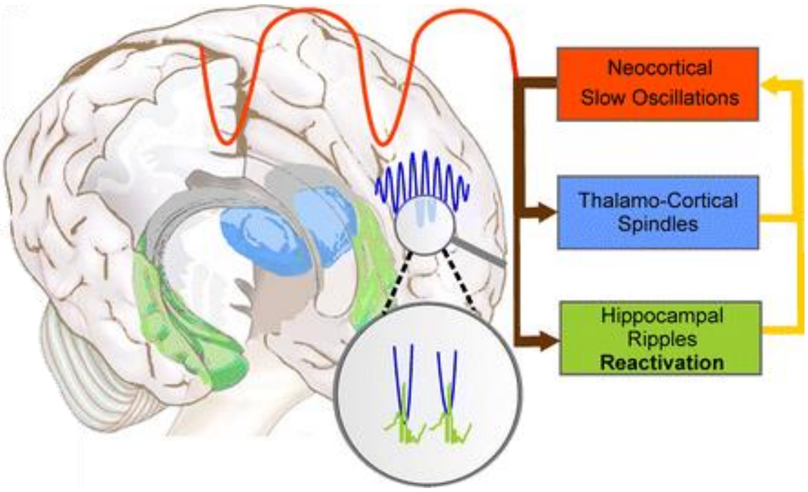

Neocortical Slow Oscillations
Thalamo-Cortical Spindles
Hippocampal Ripples Reactivation

**Figure 1. Illustrating the relationship of sleep-based neurological activity suggested to support offline consolidation.** During neocortical slow oscillations (red) thalamo-cortical spindles (blue) occur. Nested within the downstate of the spindles, hippocampal ripples can occur (green). The coordinated spindle and ripple activation, driven by neocortical slow oscillations, has been associated with reactivation and redistribution of hippocampal memory traces for long-term storage in neocortical regions (figure adapted from Rasch & Born, 2013).

#### *1.2.1.2 Linking neurological and behavioural evidence of sleep-based consolidation*

Evidence suggests that sleep is associated with the consolidation of novel words, in line with systems consolidation theory. For example, Dumay and Gaskell (2007) found clear behavioural evidence that novel words had been incorporated into the mental lexicon after sleep but not after an equivalent period of wakefulness. Participants were taught novel nonsense words in the morning (*am group*) or evening (*pm group*). They were tested immediately after the learning phase using tasks that allowed a measure of lexical competition and explicit memory of the nonsense words (e.g., *cathedruke*) that were similar to an already known word (e.g., *cathedral*). To test explicit recall, participants completed a free recall task where they were given 3 minutes to say as many of the nonsense words as they could recall. In the lexical competition test, participants made speeded decisions on whether a spoken word contained a pause (e.g., *cathedr_uke* and *cathedr_al*). The reasoning behind this test was that the reaction time in the speeded decision would be slower for the already known word once the novel nonsense word had been integrated with the mental lexicon (Gaskell & Dumay, 2003). Indeed, in agreement with their hypothesis of a lexical competition effect not emerging immediately after the learning phase, there was no effect on reaction time in the lexical competition test. 12 and 24 hours after the learning phase, participants repeated the same tests, meaning that the pm group had one period of sleep before the 12-hour test, but the am group did not sleep until the 24 hours test. If a sleep period occurred before the 12 hours test (i.e. pm group), the novel words were engaging in lexical competition with existing words (i.e., slower reaction times in the lexical competition test), and participants could recall more nonsense words in the free recall test. In contrast, if no sleep had occurred, there were no signs of lexical competition or improvements in recall performance. These findings support the role of sleep in strengthening new memories of words and actively integrating them into the mental lexicon.

Strengthening the link between EEG activity during sleep and learning is a study by Gais, Mölle, Helms and Born (2002), who found that adults that learned lists of unrelated word pairs showed increased rate and density of spindles in the following night's sleep. In contrast, adults who had not performed a learning task before sleep showed more typical rates and densities of spindles. The increased spindle densities were also positively correlated with task performance (i.e., recall of the word pairs) the day after. Similarly, it has been found that faster sleep spindle density is related to the consolidation of weakly encoded words and memories, primarily when the spindles occur in coordination with slow oscillations during SWS (Denis et al., 2021; Weighall et al., 2017; Schmidt et al., 2006). These findings suggest that spindle activity is sensitive to prior learning experiences and that increased spindle activity can benefit the consolidation of novel memories and words effectively.

Further, the semantic properties of a new word can affect the density of spindles and slow oscillation activity. For example, Tamminen, Lambon-Ralph and Lewis (2013) taught participants novel nonwords with an associated semantic concept with several or few neighbouring semantic associations. Then, using polysomnography, participants' EEG activity was recorded on the following night's sleep, allowing analysis of slow oscillation and spindle activity. Interestingly, learning novel words with few existing semantic associations lead to increased spindle activity and a trend toward higher slow oscillation activity. This finding would suggest that novel words that are harder to associate with existing knowledge require increased communication between hippocampal and neocortical areas to create new semantic representations. In other words, associating novel words with semantic properties affects later consolidation during sleep (Leach & Samuel, 2007).

It is important to note that the sections above demonstrate two different aspects of word learning; *lexical activity* (i.e. signalling how well integrated a new word is in the mental lexicon) and *explicit memory* (i.e. how many words a participant can recall). As per a systematic review and meta-analysis investigating whether sleep affects word learning differently depending on the retrieval type, there was evidence that sleep benefited performance in recall, recognition, and lexical integration (Schimke, Angwin, Cheng & Copland, 2021). While it is important to consider lexical activity when dissecting sleep-based consolidation processes, this review will mainly focus on explicit knowledge of novel words (e.g. free/cued recall, recognition). The reason for this is twofold: robust sleep benefits have been found for the explicit retrieval of newly learned words, and benefits of spaced retrieval practice (discussed later) have largely been found in the context of retrieval tasks. In their omnibus meta-analysis, Schimke et al. (2021) reviewed findings from over 1300 participants across 25 unique studies (yielding 29 separate between-subject comparisons), showing moderate effect sizes for a sleep benefit in recall and recognition memory (g=0.57 and g=0.52, respectively) and a small effect size for lexical integration tasks (g=0.39). Overall their analyses show that sleep benefits the acquisition and consolidation of novel words across a range of retrieval domains. From here on, we will mainly focus on studies using recall and recognition tasks (with some exceptions) as such methods can also be flexibly used in a broad range of testing environments (for example, in experiments taking place with children in school settings, which will be reviewed later).

#### *1.2.1.3 Sleep and encoding before consolidation*

Through the sections above, we have used existing literature to show how vital a sleep period *after* learning novel words is, as it allows offline consolidation to stabilise memory traces, aiding explicit recall. However, a period of sleep *before* learning can also prepare neural networks for further encoding. Neurological evidence for this shows that hippocampal and neocortical functioning (crucial for word learning according to the CLS account) is improved after a period of sleep (Cellini et al., 2016; Van Der Werf et al., 2009; Dolan & Fletcher, 1997; Grosvenor & Lack, 1984). Further, if sleep before learning was disrupted, for example, by suppressing slow wave activity without affecting total sleep time (Van Der Werf et al., 2009), subsequent learning can be negatively affected, enhancing the claim that sleep prepared cortical regions for learning. In other words, sleep before learning allows the selective neuronal activity in hippocampal and neocortical regions needed for effective word learning, according to the CLS accounts. Further, sleep deprivation before word learning can significantly impair memory and learning abilities (Cirelli & Tononi, 2019; see Newbury, Crowley, Rastle & Tamminen., 2021 for a meta-analysis), highlighting that sleep before (and after) learning occasions is needed for effective word learning.

Relatedly, there also appears to be a threshold of recall that memories need to reach before sleep for consolidation to have an effect (e.g., Drosopoulus et al., 2007), highlighting the importance of accounting for encoding variables before sleep. If initial recall falls below the threshold, e.g., due to a limited number of exposures, no apparent benefit from sleep-based consolidation occurs (Walker et al., 2019; Schoch, Cordi & Rasch, 2017; Drosopoulus et al., 2007). Similarly, if initial exposure levels exceed an upper threshold (e.g., 90% correct recall before sleep in Drosopoulus et al., 2007), sleep benefits appear reduced. Recently though, it has been found that consolidation processes can benefit both weakly and strongly encoded memories (Petzka et al., 2021), but if the final testing conditions are not sensitive enough, these effects could be masked. Further, memory traces of novel words that are weakly encoded before sleep have been found to benefit more from sleep-based consolidation (e.g., Denis et al., 2022; Denis et al., 2018; Cairney et al., 2016; Diekelmann, Wilhelm & Born, 2009; Drosopoulos et al., 2007), indicating that weakly encoded memories could be prioritised for consolidation during sleep.

To summarise the role of sleep in adult word learning, a period of sleep after learning novel words has been found to stabilise and, in some cases, improve recall performance due to offline consolidation processes. However, aspects prior to learning, such as preceding sleep and strength of initial encoding, can affect subsequent consolidation effects and long-term retention of novel words. In the following sections, we will explore the literature on how the structure and form of the training can affect the learning and long-term recall of novel words.

### 1.2.2 Spaced retrieval practice in adults

Two encoding methods that have produced findings with strong pedagogical and theoretical implications are *retrieval practice* (also referred to as the *testing effect*) and *spaced learning (*also referred to as the *spacing effect)*. These methods are widely regarded as promoting robust long-term retention of memories and words, particularly when combined (Cepeda et al., 2009; Cepeda et al., 2006). Studies examining retrieval practice typically have three phases; 1) an initial learning phase in which participants are exposed to the to-be-learned material, 2) a practice phase where participants are encouraged to either actively retrieve the material or passively restudy it and 3) a final test phase. When combined with a spacing paradigm, the practice phase is usually separated into shorter sessions spaced over several days or weeks (see Figure 2). Interestingly, the importance of memory consolidation during sleep has often been overlooked in theories of spaced retrieval practice (e.g. Smith & Scarf, 2017). This oversight has led to uncertainty on whether the benefits of spaced retrieval practice may be at least partly driven by sleep consolidation effects (i.e., if the sessions are spaced over periods of sleep).

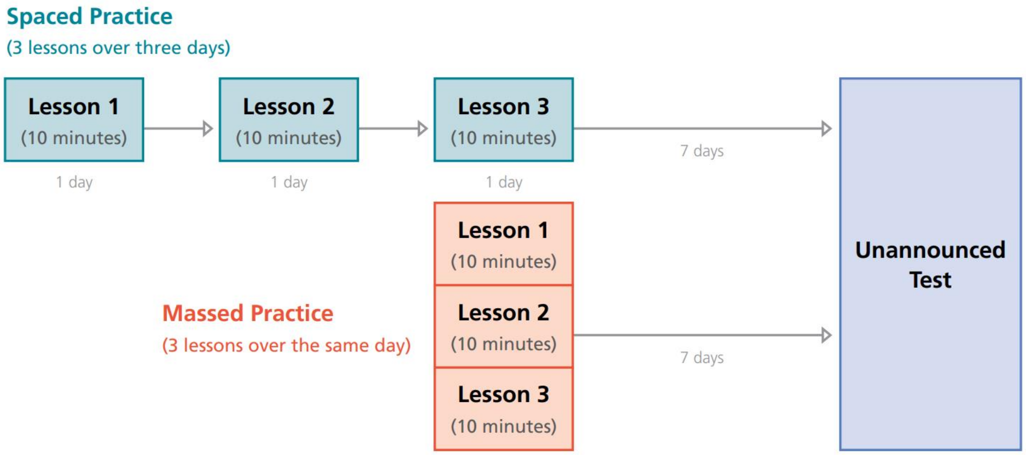


**Figure 2. A typical schedule of spaced and massed schedules of a spaced retrieval practice study.** The top illustrates three 10-minute lessons spaced 1 day apart to form a spaced retrieval practice schedule. The bottom shows three 10-minute lessons massed into one session taking place on a single day. 7 Days later, a final retention measure is acquired through a test, where a spacing benefit is often manifested as superior memory retention for participants in the spaced schedule, compared to participants in the massed schedule who tend to show increased levels of forgetting, leading to lower levels of memory retention. This schedule was utilized in Bloom and Shuell's study on spaced practice in 1981, and the figure was adapted from Carpenter and Agarwal, 2020.

The following sections will first evaluate the effects of retrieval practice and then how introducing a space between retrieval attempts (i.e., spaced retrieval practice) can affect later retention of novel words. We will mainly focus on evidence examining word and language learning in different forms (e.g., learning of word pairs, definitions of rare or complex words, nonsense words, and in rare cases, grammatical constructs). Then, we will cover common theories of why a spacing benefit may emerge in long-term memory. Following this, some questions that remain unaddressed by the theories will be outlined, including the potential role of sleep in spaced retrieval practice.

#### *1.2.2.1 Retrieval practice and the testing effect*

Retrieval practice and the testing effect (i.e., better long-term retention after retrieval practice compared to restudy) have been well-established phenomena in the memory literature for over a century (see Spitzer, 1939; Gates, 1917; and Abbott, 1909, for early investigations of benefits from testing). More recently, the interest in retrieval practice and its effects on long-term retention was reignited by a study by Roediger and Karpicke (2006). Subsequently, the literature has vastly expanded (Rowland, 2014; Rawson & Dunlosky, 2011), and the effect has been widely promoted for use in educational settings (Agarwal, D'Antonio, Roediger, McDermott & McDaniel, 2014; Roediger et al., 2011; Agarwal et al., 2009).

In Roediger and Karpicke's (2006) first experiment, students (aged 18 to 24) read two prose passages in the initial learning phase. Then, in the training phase, taking place immediately after the initial learning phase, participants reread one passage twice (restudy condition) and reread the other passage once before completing a free recall test (retrieval condition). Finally, participants completed a final retention measure (free recall test) 5 minutes, two days, or one week later. Participants tested 5 minutes after the training phase could recall the most information from the passages, and there was a small but significant benefit for the restudied prose passage. However, while the overall performance two days and one week later were lower due to forgetting, the benefit from restudying or retrieval practice was reversed, showing that the prose passage practised through free recall was better remembered than the restudied passage. Interestingly, the passage tested through free recall in the practice phase was better remembered one week later (56% correctly recalled information) than the restudied passage two days later (54% correctly recalled information). This indicates that initial retrieval practice can protect novel information more effectively over a longer period than restudying information, which showed greater forgetting after only two days. Through their experiment, Karpicke and Roediger (2006) provide clear evidence that practising retrieval of novel information significantly slows forgetting in long-term measures.

Subsequently, several studies have replicated similar findings in a range of word learning tasks, such as single word list learning (Carpenter & DeLosh, 2005) and learning of word pairs (Pyc & Rawson, 2009; Carpenter, 2009; Toppino & Cohen, 2009). In addition, several reviews and meta-analyses have been conducted (e.g., see Adesope, Trevisan & Sundararajan, 2017; Rowland, 2014 for meta-analyses; see Moreira et al., 2019; Roediger & Butler, 2011; Balota, Duchek & Logan, 2007 for reviews), which generally conclude that the testing effect after retrieval practice is robust and can emerge in several conditions.

A key finding of focus here is that a sleep period can reduce or entirely remove the positive long-term effects of retrieval practice (e.g., Abel et al., 2019; Bäuml, Holterman & Abel, 2014). In short, after a period of wake, a traditional testing benefit can be observed, but if the same period includes sleep, this difference disappears. This could potentially be attributed to the beneficial sleep effect outlined in earlier sections. In line with this, Antony and colleagues (2017) put forward a neurocognitive account suggesting that retrieval practice can act as a "fast route" to consolidation. They argue that retrieval practice can result in similar simultaneous activation of memory traces in hippocampal and neocortical regions as observed during overnight sleep. Specifically, Antony et al. (2017) suggest that retrieval practice initiated by incomplete cues (e.g. in cued recall and multiple-choice tasks) triggers coordinated hippocampal-neocortical activity. Like during sleep-based consolidation, this communication allows novel memories to integrate with existing memory traces in neocortical regions, reducing dependency on hippocampal areas. In contrast, when restudying learning material (e.g. rereading, passive exposure), the entire memory cue would be provided, resulting in less requirement for communication between hippocampal and neocortical areas. Thus, while restudying could somewhat strengthen hippocampal and neocortical memory traces, there is less need for communication between the two regions, resulting in less effective strengthening and integration of memories.

As argued in the CLS model, consolidation shifts the dependency from relying primarily on hippocampal activation to neocortical activation. Thus, according to Antony and colleagues' suggestion, the benefits of retrieval practice would be most apparent when hippocampal traces are weak, for example, after some time has passed between initial learning and the final test (allowing for the decay of hippocampal memory traces), as neocortical regions would be more engaged at later retrieval. Evidence supports this in the retrieval practice literature, where the testing effect tends to be stronger in long-term measures than in shorter-term measures (e.g., Rowland, 2014; Roediger & Butler, 2011). Further, studies exploring word learning in developmental amnesia support that retrieval practice encourages rapid integration of novel memories in neocortical areas. Developmental amnesia is when bilateral atrophies to the hippocampus cause the inability to form new memories. When participants with developmental amnesia learn new words with semantic properties in sessions involving retrieval practice, they could successfully retrieve the words later (e.g. Kim, Saberi, Wiseheart, & Rosenbaum, 2018; Green, Weston, Wiseheart, & Rosenbaum, 2014). Thus, the authors of these studies argue that retrieval practice bypassed hippocampal dependency (as the participants did not have functioning hippocampi) and arguably initiated a greater reliance on neocortical processing, in line with Antony et al.'s (2017) suggestion.

#### *1.2.2.2 Repeated vs single retrieval practice*

The sections above show that retrieval practice can potentially effectively promote faster consolidation and incorporation of novel words in the mental lexicon in typical and memory-impaired populations, enhancing long-term retention of novel words. This appears to be the case even with minimal initial exposure and no further learning opportunities (as shown in the first experiment by Roediger and Karpicke, 2006). Interestingly, when given multiple opportunities to practice retrieval, the testing effect can be enhanced, even without further exposure to the learning material (Karpicke & Roediger, 2008; Roediger & Karpicke, 2007). In a second experiment, Roediger and Karpicke (2006) examined the effect of repeated retrieval practice or restudying. The initial learning phase was the same as described above, except participants only read one prose passage once. They then utilised a between-subject design where participants were asked to (i) reread the passage three more times (*restudy condition*), (ii) reread it twice and complete one free recall test (*single retrieval practice condition*), and (iii) complete three free recall tests back-to-back (*repeated retrieval practice condition*). It is important to note that the single and repeated retrieval practice conditions did not include feedback or further exposure to the prose passage, meaning that participants in the repeated retrieval practice condition only had a single exposure to the full material. The final test phase took place immediately after the practice phase or one week later, where participants were asked to write everything they could remember from the prose passage (i.e., free recall). At the immediate test, participants in the restudy condition could recall significantly more than the two other conditions, which was not unanticipated considering the increased exposure rates. However, what is striking about this study is that one week later, the pattern had reversed and participants in the restudy condition showed greater levels of forgetting, resulting in the lowest performance of the three conditions. Participants in the repeated retrieval practice condition outperformed the other two conditions as they could recall significantly more information one week later. This reversal of performance clearly shows that, even with minimal levels of exposure, repeated retrieval practice leads to long-term benefits in recall performance. In a later study, Karpicke and Roediger (2007) showed that if a temporal delay is introduced between retrieval attempts, retention two days later can be further enhanced than if retrieval attempts occurred back-to-back (which was the case in their initial experiment in 2006). Further studies show that introducing a delay between retrieval attempts can be more beneficial than massing retrieval practice (e.g., Kornell, 2009; Carpenter & DeLosh, 2005). This effect was argued to be due to a combination of the benefits from retrieval practice and spaced learning, henceforth referred to as *spaced retrieval practice*.

To summarise the effects of retrieval practice, extensive literature reports a robust benefit in long-term measures after practising retrieval compared to restudying the same material. For example, when someone learns a novel word and practices retrieving it through a test, they will forget it at a slower pace than if they reread the same word without any retrieval effort. An interesting theory of why this occurs is in line with the CLS model, where retrieval practice has been suggested to encourage consolidation-like communication between hippocampal and neocortical regions resulting in faster integration in the mental lexicon. Compared to restudy, reduced effects from sleep-based consolidation after retrieval practice support this theory. Additionally, repeated retrieval practice occasions are more beneficial than a single retrieval practice, even without further learning opportunities during retrieval attempts. We will now examine how altering the temporal space between repeated retrieval practices can affect the long-term retention of novel words.

#### *1.2.2.3 Spaced retrieval practice*

As mentioned above, introducing a temporal delay between retrieval attempts (*spaced retrieval*) can enhance long-term retention more than if retrieval attempts occur back-to-back (*massed retrieval*). Karpicke and Roediger's 2007 study varied the delay between retrieval attempts by introducing a number of intervening items. Specifically, massed retrieval items were presented with no intervening items (e.g., *AA BB CC DD EE FF*), while spaced retrieval items were presented after five intervening items (e.g., *ABCDEF ABCDEF*). Thus, the items were presented the same number of times (2 times in the examples above), but the space between presentations varied. We will refer to this type of spacing as *spaced presentations*, as all retrieval attempts took place in a single session, but the number of intervening items between presentations was altered to create a spacing paradigm. Karpicke and Roediger (2007) found that word pairs practised through massed presentations showed an initial advantage in recall levels (i.e., they were better remembered), but after 10 minutes and two days, the advantage shifted to the spaced words being better remembered. Their finding indicated that spaced presentations of retrieval opportunities slowed forgetting in relatively long-term measures (10 minutes and two days).

Exploring the effect of spaced presentation of retrieval attempts over a longer timescale (1 week), Pyc and Rawson (2009) had adult participants learn 70 Swahili-English word pairs through repeated retrieval practice (recalling the English word when presented with the Swahili match) spaced by a short lag (approximately 1 minute) or a long lag (approximately 6 minutes). One week after initial practice, words practised in the long lag condition were consistently better recalled than those in the short lag condition. In fact, word pairs practised in the short lag condition showed a floor effect on recall one week later, regardless of how many times the participants had practised the word (participants were given up to 10 retrieval practice opportunities of each word). While this finding shows that retrieval practice separated by a 6-minute space can benefit long-term retention of word pairs one week later, it highlights that the benefit may be reduced if the space is too short (i.e., 1 minute in this case).

Similar spacing benefits can be found if retrieval practice is spaced across several days. For example, Bloom and Shuell (1981) show that a group of students who studied French words over three consecutive days could recall more words four days later than students who studied the same words in a single session on one day. Interestingly, there was no difference in the number of words recalled between the spaced and massed groups immediately after the training phase (i.e., after the massed session or after the final spaced session). Thus, the spacing benefit only emerged in the longer-term measures. This highlights an interesting pattern of the spacing effect, where the size of the spacing effect at final retrieval depends on the initial spacing delay (see Cepeda et al., 2006). In other words, to achieve the maximal spacing benefit, the space between retrieval sessions needs to increase as the delay before the final retention test increases. For example, adults learning word pairs in 2 sessions spaced seven days apart showed a stronger spacing effect (i.e., better performance) if the delay before a final cued recall test was 13 days (longest lag), as opposed to a lag of 2 or 6 days (shorter lags; Gerbier, Toppino & Koenig, 2014).

Interestingly, the window for the optimal space between retrieval attempts can be affected by the number of retrieval attempts given. As mentioned above, repeated retrieval practice can be more effective than a single retrieval occasion (Roediger & Karpicke, 2007). In the spacing literature, having a single retrieval opportunity 1 or 7 days after initial training produced 86% and 62% recall levels after one week (Küpper-Tetzel, Kapler & Weiseheart., 2014). In contrast, if four retrieval opportunities were given three or 14 days after initial exposure, recall levels were 83% and 81%, respectively, after a 7-day delay (Bird, 2011). This shows that when multiple retrieval opportunities were given, varying the initial space between retrieval practices did not impact retention levels one week later. This section highlights that the spacing effect can be affected by variables such as the number of retrieval attempts or initial spacing relative to the final retention interval. Before delving deeper into what other aspects could impact the spacing effect, we will outline current theoretical accounts of why a spacing effect emerges in the first place.

### 1.2.3 Why is spaced retrieval practice beneficial for long-term retention in adults?

#### *1.2.3.1 Contextual variability*

Contextual variability theories provide one account of why spaced retrieval practice enhances long-term retention (Estes, 1955; Pashler et al., 2009; Maddox, 2016). According to these theories, spaced retrieval practice allows for a greater variety of contextual cues to be integrated with novel memory traces, creating more cues to aid long-term retrieval and leading to better longer-term memory performance. Importantly though, these theories predict an immediate benefit from massed retrieval practice (i.e., measured at the end of the practice period) as the similar context at each retrieval occasion would mean more overlapping cues aiding immediate retrieval. In contrast, when retrieval practice is spaced over intervening items or time, the context of each retrieval occasion would vary, eventually resulting in fewer overlapping cues available for irecall but a greater variety of cues to aid retrieval in longer-term tests. Indeed, extant literature suggests an immediate benefit from massed retrieval practice in adults, especially in language learning tasks and a later emerging benefit from spaced retrieval practice in long-term retention (e.g., Suzuki & DeKeyser, 2015; Simone, Bell & Cepeda, 2013; Delaney, Verkoeijen & Spirgel, 2010).

#### *1.2.3.2 Desirable difficulty/Retrieval effort*

Another explanation for why spaced retrieval practice yields better long-term retention than massed retrieval practice is because the temporal space between retrieval attempts allow forgetting to occur, meaning that the subsequent spaced retrieval attempt meets the *desirable* *difficulty* (e.g., Schmidt & Bjork, 1992; Bjork & Bjork, 2011). A retrieval attempt meets the desirable difficulty when the retrieval is hard, but not so hard it prevents correct recall. Pyc and Rawson (2009) built on this to propose the *retrieval effort hypothesis*, arguing that difficult retrieval leads to better long-term retention (than easier retrieval) because more challenging retrieval activates memory traces that decay slower (Pavlik & Anderson, 2008). The slower decay of the memory traces results in slower forgetting compared to faster decaying memory traces from easy retrieval attempts, eventually leading to the spacing benefit observed in the literature. Crucially, Pyc and Rawson's experiments demonstrated that longer intervals between retrieval attempts (i.e. spaced retrieval practice) led to enhanced recall levels than when retrieval attempts were not separated in time (i.e. massed retrieval practice).

Further support for this is a pattern of diminishing returns of benefits in later recall after three successful retrieval attempts provided by Rawson and Dunlowsky (2011). According to their findings, the effect on later retention was reduced when retrieval was easy enough to lead to three successful retrieval attempts. Corresponding with the retrieval effort account, retrieval attempts spaced out over time require more effort than if they occur close together (owing to forgetting over time), explaining why spaced retrieval practice leads to better long-term retention than massed retrieval. The retrieval effort hypothesis can also explain the immediate benefits of massed retrieval practice in adults (i.e., easier retrieval leads to better immediate performance following massed retrieval practice but faster forgetting and lower performance in long-term measures).

A caveat to the retrieval effort theories is that it remains unclear how sleep and retrieval effort interact. For example, a behavioural indication of harder retrieval is lower levels of recall, which could elicit a greater effect from sleep-based consolidation. Thus, when studies examine spaced retrieval practice and find an initial disadvantage from spaced practice (i.e., lower initial performance) followed by better long-term recall levels, we cannot distinguish if this is an effect of retrieval difficulty or enhanced consolidation processes due to the lower levels of initial retrieval.

#### *1.2.3.3 Reconsolidation explanation*

While there were early attempts to consider the role of sleep in explaining the benefits observed from spaced retrieval practice (e.g. Landauer, 1969), subsequent accounts rejected this idea due to methodological weaknesses (see Delaney et al., 2010 for a review). However, recent advancements in our understanding of sleep-based consolidation highlight the importance of including consolidation in theoretical accounts of learning methods such as spaced retrieval practice. Thus, Smith and Scarf (2017) put forward a *reconsolidation* account of spaced retrieval practice.

Reconsolidation is when a partially consolidated memory trace is reactivated (e.g. through retrieval at a later stage), placing the memory trace in a fragile state, allowing additional cues and contexts to be integrated with future consolidation (e.g. during sleep). Studies using protein synthesis inhibitors in rodents have found that reconsolidation affects the strength of memory traces and can slow down forgetting (e.g. Lee et al., 2005; Lee, 2008; see Alberini, 2011 for a review). Similar discoveries have been made in humans, where reconsolidation strengthens and alters original declarative memory traces to include new knowledge and experiences (e.g. Coccoz, Maldonado & DeLorenzi, 2011; Forcato et al., 2007). Therefore, it could be suggested that reconsolidation plays a role in both retrieval practice (i.e. retrieval alone can cause reactivation of memory traces, Antony et al., 2017) and spaced retrieval practice (i.e. partially consolidated memories are reactivated in spaced sessions).

In human word learning, evidence of a different neural mechanism for retrieval with or without a space comes from Vilberg and Davachi (2013). In their study, participants practised word-object pairs in sessions separated 20 minutes (*massed*) or 24 hours (*spaced*) apart. While practising the word-object pairs, neural activity was recorded using fMRI. It was found that word-object pairs practised in spaced sessions elicited greater connectivity between hippocampal and neocortical regions than word-object pairs practised in the massed session. Furthermore, this connectivity was negatively associated with the forgetting of word-object pairs in the spaced but not massed sessions. In other words, stronger connectivity resulted in less forgetting after spaced retrieval practice, but not after massed. These findings would support the argument that providing a space between practices enhances the effectiveness of subsequent retrieval practice due to reconsolidation processes (i.e. supporting Smith and Scarf's account).

The spacing literature contains a wide range of lengths of spaces. For example, retrieval practice attempts can be spaced through interleaved presentations (Karpicke & Roediger, 2007), minutes (Pyc & Rawson, 2009), days (Küpper-Tetzel et al., 2014; Bird, 2011), or months (e.g., Cepeda et al., 2009). When considering the reconsolidation account, the range of spaces causes an issue, as the above examples are all considered spaced retrieval practice, but some studies contain intervening sleep between retrieval attempts, and others do not. In addition to this potentially being an issue when considering the reconsolidation account, certain studies report findings that align with the sleep effects of offline consolidation. Thus, it is vital to distinguish between sleep and no-sleep spaced retrieval practice and attempt to distinguish between the spacing effect and sleep effect. The following section will further evaluate sleep's potential role in spaced retrieval practice.

### 1.2.4 Sleep's role in spaced retrieval practice

As mentioned, it is only recently that sleep has been considered a viable variable that may contribute to the long-term benefits of spaced retrieval practice. Indeed, many retrieval practice studies span several days and thus several sleep opportunities. In such studies, it is impossible to separate whether the beneficial effects come from spaces or sleep. Therefore, a way to determine whether spaced sessions are more effective than massed sessions without the influence of sleep is by removing the intervening sleep period between spaced retrieval attempts, for example, through within-day spaced retrieval practice.

Few studies have examined within-day spaced retrieval practice where the space between retrieval attempts spans hours rather than seconds or minutes (i.e., spaced *sessions*, not spaced *presentations*). One example is Bell, Kwadri, Simone and Wiseheart (2014), who examined spaced retrieval practice spanning a day of wakefulness or over a night of sleep and compared these conditions to a single massed session. Undergraduate students in this study learned Swahili – English word pairs and practised retrieving these in two sessions on different schedules. The schedules were either *massed* (both sessions taking place back-to-back), 12 hours apart on the *same day* (e.g. 9 am and 9 pm), 12 hours including *sleep* (e.g. 9 pm and 9 am), or *24 hours* apart (e.g. 9 am on two consecutive days). Ten days later, participants returned for a final cued recall measure of long-term retention. A clear pattern emerged at the end of the second retrieval practice session: participants who practised spaced retrieval practice over a sleep period could recall significantly more than the groups that did not sleep. This initial finding would indicate that a sleep period would be crucial for an early emerging spacing effect. Interestingly, however, the performance difference between the 12-hour sleep and no-sleep groups disappeared at the 10-day test, suggesting that sleeping between spaced sessions did not exert a long-term benefit. However, the 24-hour spaced group still performed higher than all the other groups in their long-term measures. Thus, their findings provide preliminary evidence that sleep between retrieval practice sessions spaced 12 hours apart did not influence the retrieval practice spaced benefits ten days later, but a 24-hour space was beneficial.

As the participants in Bell et al.'s (2014) study only completed one additional retrieval session after the initial training, it is possible that a space of 24 hours was within the optimal spacing window for promoting long-term retention, but a 12-hour space was too short regardless of whether sleep was included or not. Further, this study was limited as there were baseline differences between the groups at the immediate test, complicating interpretation. That is, the within-day spaced group (12-hour same-day group) performed lower than the across-days spaced groups (12-hour sleep group and 24-hour spaced group) before the long-term measures. This difference could be interpreted as retrieval being harder in the within-day spaced group (due to lower performance), which according to the retrieval effort hypothesis, should result in slower forgetting. Compared to the 12-hour sleep group, this appears to be the case as the difference between these groups was not significant ten days later, but it cannot account for the higher performance in the 24-hour group as they already performed higher before the long-term measure (potentially suggesting easier retrieval which should result in more forgetting in long-term measures; Pyc & Rawson, 2009). Alternatively, the initially better recall performance for the 12- and 24-hour sleep group could be due to sleep-based consolidation processes and reconsolidation enhancing performance after sleep. Similarly, the 12-hour same-day group could have benefited from subsequent consolidation opportunities stabilising memory traces, reducing the difference between the two 12-hour groups. To explain the full findings of Bell et al. (2014), we can therefore consider a combination of theories; retrieval effort theories and sleep effects could explain performance at the immediate measure, but the retrieval effort theories cannot explain the lack of forgetting in the 24-hour group in long-term measures, while the reconsolidation account can. Thus, in short-term measures and the immediate effects of spaced retrieval practice, it seems that more than one theory can explain the performance pattern regardless of whether a sleep period occurred, but the reconsolidation account can account for changes in long-term measures.

More recently, Kroneisen and Kuepper-Tetzel (2021) explored whether a spacing benefit emerges when retrieval practice is spaced across more extended periods. In an initial training session, young adults were exposed to Lithuanian-English word translations. Immediately after or after a 2-hour delay, participants practised retrieving the English word after being presented with the Lithuanian word as a cue. Importantly, participants were not provided feedback, so no additional encoding opportunities were provided after the initial training. Then, 12 hours later, a final test was delivered after a period of wake or sleep. In this test, participants who practised retrieving the words immediately after learning and then underwent 12 hours of wake showed steeper forgetting than those who slept during the 12 hours. Interestingly, however, when the words were practised after a 2-hour delay, they did not show forgetting, regardless of whether the following 12-hour period included sleep or wake. The authors argued that delaying retrieval practice by 2 hours led to more robust memory representations than if the test was immediate. This suggests that spaced retrieval practice (without intervening sleep) is beneficial relative to massed retrieval practice and that subsequent sleep-associated consolidation effects can be attenuated by post-learning spaced retrieval practice. A possible explanation for this could be that retrieval practice acts as a form of consolidation (Antony et al., 2017). Specifically, due to the short time between retrieval attempts in the massed condition, retrieval could rely more on local short-term memory processes than coordinated communication between cortical regions (which is argued to be required for effective retrieval attempts, Antony et al., 2017). In other words, the retrieval in the massed condition was insufficient to elicit consolidation processes, resulting in forgetting if no sleep-based consolidation occurred immediately after. Instead, when practising retrieval after a 2-hour interval, the novel words underwent a consolidation process, resulting in less forgetting at later tests. Kroneisen and Kuepper-Tetzel (2021) show that within-day spaced retrieval practice (without feedback) can have similar effects as a period of sleep (i.e., offline consolidation), but as they did not include a long-term measure, it remains uncertain if this effect maintains later.

The above studies have shown that within-day spaced retrieval practice can effectively promote memory retention in certain contexts (e.g., Kroneisen & Kuepper-Tetzel, 2021), but in other cases, they are overshadowed by the potential interference of sleep effects (e.g., Bell et al., 2014). However, the studies above fail to address certain questions. For example, Kroneisen and Kuepper-Tetzel (2021) provided an extensive training procedure resulting in high initial recall performance; thus, their findings were susceptible to ceiling effects. This leaves the question of whether spaced retrieval practice can effectively promote further improvements in performance if the initial training paradigm is not as extensive, resulting in lower initial encoding levels. Further, both Bell et al. (2014) and Kroneisen and Kuepper-Tetzel (2021) only included one additional round of retrieval practice after initial training, and as previously outlined, the amount of retrieval practice can affect the window of optimal effectiveness from a spacing schedule (Küpper-Tetzel et al., 2014; Bird, 2011). Thus, we could hypothesise that the benefits from within-day spaced retrieval practice would be greater in both immediate and long-term measures if the participants in the above studies were given more retrieval opportunities.

Finally, a key difference between Bell et al. (2014) and Kroneisen and Kuepper-Tetzel (2021) is that Bell et al. provided feedback in the retrieval attempts, while Kroneisen and Kuepper-Tetzel did not. This difference is critical, as providing feedback would provide participants with an additional learning opportunity (akin to spaced learning), potentially affecting later retention of novel words differently than if no feedback was provided. In the next section, we will examine the role of feedback in spaced retrieval practice closer.

### 1.2.5 Feedback and spaced retrieval in adults learning novel words

Many studies examining spaced retrieval practice include corrective feedback in retrieval attempts. Including corrective feedback with retrieval attempts can have many beneficial effects, such as improved performance (Vojdanoska et al., 2010; Butler, Karpicke & Roediger, 2007) and enhanced motivation to continue learning (Abel & Bäuml, 2020). However, an important caveat to including feedback in retrieval attempts is that participants are exposed to the learning material multiple times, reducing the distinction between spaced *retrieval practice* and spaced *learning*. Thus, it is important to consider whether the inclusion or exclusion of feedback may influence the long-term effects of spaced versus massed retrieval practice.

#### *1.2.5.1 Feedback's effects on retrieval practice*

First, it has been found that including feedback can affect the effectiveness of retrieval practice. The findings of the literature on retrieval practice with and without feedback is mixed, where studies report both enhanced effects from retrieval practice when feedback is included (e.g., Kang, McDermott & Roediger, 2007; Karpicke & Roediger, 2007; Pashler et al., 2005), while some studies report reduced effects from spaced retrieval practice when feedback is introduced (e.g., Pastötter & Bäuml, 2016).

For example, Pastötter and Bäuml (2016) explored whether including feedback at the final test phase would affect subsequent performance on additional tests depending on if initial learning was focused on retrieval practice or restudy. Participants learned loosely related word pairs (e.g. towel – linen) and practised retrieving or restudying them twice on the first day of the study. Two days later, participants came back and completed two cued recall tests (e.g. towel – _____). Participants received feedback on the correct word pair after each trial in the first cued recall test. The second cued recall was without feedback and functioned to measure retrieval levels. They found higher recall levels on the first test for word pairs initially practised through retrieval than restudy. Interestingly, however, in the second cued recall test (i.e. after receiving feedback), initially restudied words were better recalled than retrieved words, effectively reversing the earlier observed benefit from retrieval practice. Their findings suggest that including feedback in the final test phase eliminates the benefits of retrieval practice and that restudying with feedback leads to better recall two days after learning. To explain their findings, they assume that the restudied words were close to the threshold needed for successful retrieval at the first test but did not reach the threshold. Feedback then pushed the restudied words over the threshold for successful retrieval, resulting in improved performance in the second test, removing the initial gap between words practised through retrieval. However, this study did not include further recall measures after a longer interval (e.g., one week later), so we do not know if this revered pattern was maintained in the longer term.

Further, Abel and colleagues (2019) explored the effects of feedback in retrieval attempts on subsequent sleep effects. As outlined previously, retrieval practice can lead to smaller sleep effects than restudying (e.g., Abel et al., 2019; Bäuml, Holterman & Abel, 2014), potentially because retrieval practice can act as a fast route to consolidation (Antony et al., 2017). In Abel et al.'s (2019) study, young adults practised retrieval or restudied paired associates with or without corrective feedback in a single session. Twelve hours later (spanning sleep or wake), they completed a retention test allowing the authors to examine the effects of retrieval practice or restudy with and without feedback on retention after wake or sleep. Here, Abel et al. report a benefit of sleep on recall after both restudy and retrieval practice with feedback, but if feedback was not included in retrieval practice attempts, there were no benefits after sleep. This could indicate that retrieval practices without feedback were sufficiently consolidated (Antony et al., 2017) to not benefit from further offline consolidation. However, when feedback was included in retrieval attempts, the additional exposure to the stimuli encouraged further learning, which had to be consolidated during sleep. Thus, including feedback in retrieval practice can encourage continuous learning, which benefits from offline consolidation, explaining the findings of Abel et al. (2019). Further, this indicates a relationship between retrieval practice, feedback, and sleep, highlighting that feedback and delays, including sleep (e.g., spaced retrieval practice), could affect the long-term retention of novel words in spaced retrieval practice, which we will explore further below.

#### *1.2.5.2 Feedback's effect on spaced retrieval practice*

While few studies have explored the effects of including or exploding feedback in spaced retrieval practice (see Chapter 2 for an outline), several studies have explored the effects of delayed feedback (e.g., Guo, 2018; Carpenter & Vul, 2011; Smith & Kimball, 2010; Agarwal et al., 2009; Butler, Karpicke & Roediger, 2007). For example, Smith and Kimball (2010) conducted a study which investigated the learning of trivia facts, practised through cued recall at alternating schedules. Participants completed one round of cued recall practice 8-10 (*short lag*) or 16-20 minutes (*long lag*) after initial exposure to the trivia facts. Feedback (correct answers) was then provided immediately after the retrieval round (*immediate feedback*) or 8 minutes after the first cued recall trial (*delayed feedback*). One week later, participants completed a final cued recall test, where participants who practised retrieval after a short lag showed a higher level of correct recall if feedback was delayed. In contrast, if retrieval practice took place after a longer lag, there was no difference in retention depending on whether feedback was presented immediately after or after a delay.

More recently, Smith and Kimball's findings were replicated by Guo (2018), using words in a foreign language as stimuli and longer lags between initial exposure and retrieval practice. In their study, participants were exposed to foreign words and their definitions in an initial exposure session followed by a retrieval practice session one (*short lag*) or three days (*long lag*) later, where participants completed a cued recall test (presented with a definition and asked to write down the corresponding word). Feedback (correct answers) was provided immediately after the retrieval practice session (*immediate*) or the day after (*delayed*). Then, nine days after the retrieval practice session, participants completed a final retention test. Like Smith and Kimball (2010), Guo found that participants who practised retrieval after a short lag could recall more words nine days later if feedback was provided after a delay. Further, they also report that all participants who received delayed feedback could recall more words in their long-term measure than participants who received immediate feedback. Therefore, these two studies demonstrate that delaying feedback by minutes (Smith & Kimball, 2010) or days (Guo, 2018) enhances long-term retention of novel information if retrieval practice occurs immediately after exposure (Carpenter & Vul, 2011; Agarwal et al., 2009; Butler, Karpicke & Roediger, 2007).

However, a methodological issue with the feedback-delay studies above is that feedback in the delayed conditions was presented as a separate session, essentially providing a spaced learning session. Thus, the authors of the above papers attribute their findings to being because of the spacing effect. In other words, they argue that delayed feedback provides a new opportunity to encode the material, leading to enhanced retention (Guo, 2018; Butler, Karpicke & Roediger, 2007). Further, Guo's findings support the reconsolidation account of spaced retrieval practice (Smith & Scarf, 2017). Their findings could be interpreted such that participants could integrate the feedback with the previous retrieval attempt, indicating that the original memory trace was placed in a fragile state at delayed retrieval, allowing feedback to alter it (i.e., correcting it) before further offline consolidation. However, the question of whether feedback presented in the same session as spaced or massed retrieval practice effectively enhances long-term retention remains, especially if no sleep occurs between retrieval attempts.

### 1.2.6 Interim summary

Based on the literature reviewed in the above sections, spacing retrieval practice can lead to better long-term retention of novel information and memories (e.g., Küpper-Tetzel et al., 2014; Pyc & Rawson, 2009; Cepeda et al., 2009, 2006; Karpicke & Roediger, 2007; Bloom & Shuell, 1981); however, some studies provide conflicting evidence (e.g., Gerbier, Toppino & Koenig, 2015; Bird, 2010). Furthermore, the role of sleep in accounting for reported spacing effects has been overlooked. To address this, studies comparing within-day spacing to massed retrieval practice (thereby removing the sleep period between spaced retrieval attempts) and testing at the end of the first day and a long-term follow-up will be critical in allowing us to examine whether there are immediate and long-term benefits from spacing when sleep does not occur between retrieval attempts. Another important question to address is whether the act of spaced retrieval practice itself is sufficient to enhance later recall or whether the feedback is required to boost such effects. Addressing these questions will be necessary for advancing theories such as the reconsolidation account and, pedagogically, for informing educators on how to bolster the long-term retention of new material.

## 1.3 Word learning and spaced retrieval practice in children

### 1.3.1 Word learning in children

Like in adults, when children learn novel words, they must be incorporated with the mental lexicon to allow long-term retention and efficient use of the words in everyday language use. Unfortunately, many of the dominant models of memory lack a developmental perspective. Relevant to this is the CLS account (O'Reilly et al., 2014; McClelland, McNaughton & O'Reilly, 1995), where it is commonly assumed that the mechanisms underlying learning and consolidation remain stable over development (e.g. McClelland, McNaughton & Lampinen, 2020; Norbury, Griffiths & Nation, 2010). As outlined above, the CLS model applied to word learning in adults (Davis & Gaskell, 2009) places great value on the role of sleep in consolidation. Thus, if the same model can be applied to children, we would expect to see a similar sleep effect as adults do. Below we will outline studies examining the effect of sleep on word learning in children to determine whether models like CLS can be applied in developmental populations.

A seminal study by Henderson, Weighall, Brown and Gaskell (2012) explored how a period of wake or sleep affected the retention and lexical activity of novel words. They taught children aged 7-12 novel nonwords (e.g., *biscal*) and tested explicit memory using a recognition and cued recall test immediately, after 12 and 24 hours, and one week later. A crucial point of their methodology was that the 12-hour period contained either a period of sleep or a period of wake, allowing the authors to examine the effects sleep has on the retention of novel words. Their findings were striking in that children tested after 12 hours including sleep, showed significant performance improvements in both recall and recognition tests (and the emergence of lexical competition, which is not considered here). However, children who did not sleep during the 12 hours did not improve their performance until after at least one sleep period (i.e., in the 24-hour test). Further, one week later, all children showed slight improvements in recall performance. In short, their findings indicate that, like in adults (Dumay & Gaskell, 2007), recall of newly learned words improved after at least one period of sleep had occurred (i.e., after 12 hours if a sleep period was included, but not until 24 hours later if the initial 12 hours period spans across wakefulness), indicating that the CLS model could be applied to word learning in children. Similar patterns have been reported in a range of contexts, such as storybook learning (Williams & Horst, 2014), learning words explicitly with meaning (Henderson, Weighall & Gaskell, 2015), on a variety of lexical competition tasks (Weighall et al., 2017; Henderson, Weighall, Brown & Gaskell, 2013), and when learning word-object pairs (Horvath et al., 2015) and verbs (Xiaoxue He et al., 2020), consistently highlighting the importance of sleep in children’s word learning.

Further highlighting how sleep can aid word learning in children are findings of similar correlations between overnight changes in word memory and aspects of sleep architecture in children. As outlined in Section 1.2.1, sleep spindles occurring during SWS have been found to correlate with behavioural changes in word learning (e.g., Tamminen et al., 2010). Studies with children as young as 3, and as old as 13, have shown that the number and intensity of sleep spindles positively correlate with improvements in recall and recognition performance, both during daytime naps (Spano et al., 2018; Kurdziel, Duclos & Spencer, 2013) and overnight sleep (Smith et al., 2018; Hoedlmoser et al., 2014). Thus, like adults, offline consolidation during sleep is supported by spindles in childhood.

The sections above briefly show how sleep-based processes support word learning during childhood (see Chapter 3 for further evidence). However, and importantly, while there are apparent similarities between sleep’s benefit for word learning in adults and children (i.e., both adults and children benefit from a sleep period more than a period of wake), closer inspection of the literature reveals subtle differences. For example, children sleep for longer and have shown a greater proportion of SWS and SOs than adults (Ohayon et al., 2004), potentially affecting the effectiveness of sleep-based consolidation.

#### *1.3.1.1 How word learning differs from childhood to adulthood*

A critical question is how mechanisms of learning and retention differ over development. This is particularly pertinent in the area of language learning as children have long been known to surpass adult's ability to learn in language-related tasks (e.g., Newport, 1990), despite adults showing superior cognitive abilities in several areas related to language learning (Steber & Rossi, 2021; Craik & Bialystok, 2006). Thus, we will mainly focus on the role of sleep in word and language learning to determine whether this may be linked to children’s superior language learning abilities.

As mentioned, children have been found to show a greater proportion of SWS and SO (around 40% of total sleep time) than adults (around 20% of total sleep time; Wilhelm et al., 2013), which appears to peak at around 10-12 years of age and then decline through adolescence until it reaches adulthood levels (Carskadon, Acebo & Jenni 2004; Ohayon et al., 2004). Further, there is tentative evidence that the number and intensity of sleep spindles reach a peak in the first decade of life (Kurth et al., 2010) which then declines through adolescence (Nicolas, Petit, Rompre, & Montplaisir, 2001). In short, children generally show increased levels (compared to adults) of sleep-based processes that have been linked to offline consolidation. However, this leaves a question of whether these increased levels of consolidation-related activities during sleep in children translate to increased performance in behavioural measures.

Indeed, there is accumulating evidence that children show greater sleep effects than adults in language learning tasks (e.g., James, Gaskell & Henderson, 2019; Weighall et al., 2017; Wilhelm et al., 2013). For example, Weighall et al. (2017) showed a larger overnight benefit in recalling novel words for children over adults. Young adults and children aged 7 to 9 were taught word-object pairs and tested on recall and a visual selection task after a period of wake or sleep. To test recall of the word-object pairs, participants were asked to recall the paired word after being provided with the first part of the word-object pair. Children could recall 36% more words if they were given an opportunity for offline consolidation than if tested after wake. While adults also showed greater recall levels when given a period of consolidation, this benefit was significantly smaller (24%) compared to the benefit in children. Further, in the visual selection task, participants were presented with four pictures of objects and asked to select a target while their fixation time was recorded. Each presentation included a target object, one newly learned object, and two distractors. Both adults and children showed a longer fixation time on the learned object before making their selection, indicating that the newly learned objects competed with the target object for activation. However, only children showed a significantly greater increase in fixation time after a period of sleep, indicating that they benefited from overnight consolidation. These behavioural findings fit with the above outlined developmental differences in sleep architecture, as children showed greater benefits from a period of sleep than adults.

Another critical point is that children have continuously improved memory recall from multiple nights of sleep without further practice (e.g. James, Gaskell & Henderson, 2019). For example, children who learned novel words showed higher levels of recall after one night's sleep, and recall was further improved one week later. In contrast, adults in the same study maintained similar levels of recall after one night's sleep and one week later (James et al., 2019). This could suggest that children can more effectively consolidate novel words over several nights, leading to improved recall, while adults rely on sleep to reduce forgetting and maintain performance. Thus, strengthening the argument that children rely on sleep-based consolidation to a greater extent than adults.

However, whilst some studies report enhanced consolidation effects for children relative to adults (e.g., Henderson et al., 2015; Wilhelm et al., 2013), it is important to note that there are also studies that report equal sleep benefits for adults and children (Henderson, Weighall, Brown & Gaskell, 2013; Wilhelm, Diekelmann & Born, 2008). For example, children aged 6-8 showed twice as much SWS during overnight sleep compared to adults after learning semantically associated word pairs, but this difference was not reflected in behavioural measures (Wilhelm et al., 2008). Thus, there may be additional aspects, apart from sleep-based consolidation processes, affecting the retention of novel words in adults and children.

One variable potentially explaining the differences in word learning between adults and children could be prior knowledge, which has been associated with overnight consolidation effects (James, Gaskell & Henderson, 2019). Adults generally know more words than children, thus having a more extensive mental lexicon to aid novel word learning. For example, adults can use prior semantic and phonological knowledge to learn new words effectively without a period of offline consolidation (James, Gaskell & Henderson, 2019). In contrast, children have less prior knowledge and thus show slower learning before sleep but catch up with adults' performance after a week of consolidation (James, Gaskell & Henderson, 2019). This, therefore, accounts for why adults can outperform children initially, but children may undergo greater strengthening of novel words over a consolidation period that includes sleep.

Because of the potential influence of prior knowledge on word learning in adults and children, an important consideration when designing word learning studies to compare adults and children directly is to ensure the levels of learning are appropriate for the age. For example, adults can generally learn more words in a short time, potentially due to their more extensive prior knowledge (e.g., Steber & Rossi, 2021; Cepeda et al., 2006; Craik & Bialystok, 2006; Bahrick & Hall, 2005), meaning that they could reach a ceiling for further improvements earlier than children. In contrast, children have a smaller capacity to learn many words without an intervening period of offline consolidation (e.g., Henderson et al., 2012; Childers & Tomasello, 2005), potentially leaving more room for improvements over sleep periods. Thus, we must continue to consider how encoding and prior knowledge aspects can affect word learning in children.

### 1.3.2 Spaced retrieval practice in children

While the literature on the spacing effect is expansive (over 1000 published studies), only a small proportion of these explore the effect of spaced retrieval practice in children (see Cepeda et al., 2006; Knabe & Vlach, 2020, for reviews), and to our knowledge, none have directly compared adults and children at the time of writing. This empirical gap is surprising, particularly given that spaced retrieval practice has been put forward as a "neuro hit" in educational settings (Kang, 2016; Son & Simon, 2012) and because employing spaced retrieval practice in classrooms can help reduce test anxiety in middle and high school students (Agarwal, D'Antonio, Roediger, McDermott & McDaniel, 2014).

An early example of a study on spaced retrieval practice in older children is Bloom and Shuell (1981). They taught high-school children (aged 13-15) French words who practised retrieving these in one 30-minute session on one day (*massed*) or three 10-minute sessions on three consecutive days (*spaced*). When tested on final recall levels immediately after the final practice session, both groups of children could recall similar levels of words. However, when tested four days later, the spaced group could recall significantly more words (35% more words) than the massed group, showing a benefit from spaced retrieval practice.

An important aspect of spaced retrieval practice in children is the environment in which the training takes place. Most of a child's learning experience occurs in the classroom (i.e., in a group setting), meaning we must understand whether spaced retrieval practice benefits word learning outside a laboratory setting. Indeed, classroom-based studies have reported spacing benefits in the retention of newly learned words (e.g., Goossens et al., 2016; Goossens et al., 2012; Sobel, Cepeda & Kapler, 2011). For example, Sobel, Cepeda and Kapler (2011) taught 46 children (10 years old) eight novel words and their definitions. Participants then completed what the authors describe as a spaced learning session one minute (massed) or one week (spaced) later. Despite the authors' use of the term spaced learning, the sessions contained retrieval practice tasks (e.g., cued recall) in combination with presentations of the word-definition pairs. Therefore, this design can be interpreted as spaced retrieval practice with feedback. Five weeks later, participants who practised the novel words in the one-week spaced session could recall 177% more words than if practised in the massed session, showing a substantial spacing benefit.

However, an important finding in the spacing literature on children is that a spacing benefit can emerge immediately after the practice phase. For example, when learning novel words, children as young as two years old have shown an immediate benefit (i.e. better recall after the final practice session) from the spaced practice of novel words (Childers & Tomasello, 2002). Moreover, this spacing benefit maintains one day and one week after, suggesting that spaced retrieval practice benefits both immediate and long-term recall of novel words in younger children. Similar results have been found in word learning tasks in preschool children (e.g. Haebig et al., 2019; Leonard et al., 2019; Fritz, Morris, Nolan & Singleton, 2007) and primary school children (e.g. Petersen-Brown et al., 2019; Goossens et al., 2012; these will be further outlined in Chapter 3). In contrast, while adults can show an early emerging spacing benefit (Bell et al., 2014), a commonly occurring pattern is of equal performance (Kroneisen & Kuepper-Tetzel, 2021; Karpicke & Roediger, 2007; Bloom & Shuell, 1981) or an initial massed benefit (e.g., Roediger & Karpicke, 2006). We will now examine these subtle developmental differences in more detail below.

#### *1.3.2.1 Comparing spacing benefits between adults and children*

Metcalfe, Kornell & Finn (2009) examined whether immediate or delayed feedback led to different benefits in word learning for children and adults. Since delayed feedback can act as a spaced opportunity to strengthen memory traces following retrieval attempts, this study can be used to inform potential developmental differences in spaced retrieval practice. In an initial learning phase, school-aged children (aged 10-12) and undergraduate students were taught uncommon English words and their definitions. They then completed one retrieval practice session, with immediate corrective feedback presented after the retrieval practice round, delayed corrective feedback (provided the day after), or no feedback. The immediate feedback condition can be likened to a massed practice session (as both retrieval and feedback were presented in the same session), whereas the delayed feedback condition is akin to a spaced practice session (as retrieval and feedback were separated by one day). Participants completed a test to measure long-term recall one day after the final practice session. The proportion of novel words correctly recalled in the final test session was higher for words with delayed feedback than words with immediate feedback for both adults and children (both were higher than no feedback). Beyond lending support to a spacing benefit, these results suggest that adults and children benefit from delayed (spaced) feedback relative to immediate (massed) feedback. Metcalfe et al. (2009) ran additional analyses to control for the time that passed between the final feedback occurrence and the final test. This was important because the presentation of delayed feedback was closer to the final test than the immediate feedback. Interestingly, when controlling for the total delay before the final test, children still showed a benefit in the retention of words after spaced feedback, but the benefit disappeared in adults. This suggests that children benefited more from delayed (or spaced) feedback than adults, who benefited equally from immediate and delayed feedback.

The findings from Metcalf, Kornell and Finn (2009) mimic the pattern of results in the spaced retrieval practice literature where children benefit from spaced retrieval practice at an earlier stage (i.e. only one day had passed between final practice and final test) than adults. Based on the extant spacing literature, it could be speculated that a benefit from delayed feedback could have emerged for adults if tested again after an extended time (e.g. one week). However, it is important to mention that this study did not directly examine spaced or massed retrieval practice but focused on the effects of feedback timing. So while this study indicates that adults and children benefit from immediate and delayed feedback differently, preventing direct implications from being drawn about spaced versus massed retrieval practice.

More recently, Vlach, Bredemann and Kraft (2019) conducted a study exploring differences in metacognitive biases regarding the effectiveness of learning from spaced versus massed presentations in adults and children. Adults and children (aged 2-10, n=102 across two experiments) learned picture pairs in massed presentations (the same pair presented three times in a row) or spaced presentations (each pair presented with five intervening pairs). After a 3-minute delay, participants were tested on their associative memory of the picture pairs via a cued recall task (i.e. participants saw the first picture of a pair and three options and were tasked with choosing which of the three belonged to the first picture). Here, children consistently showed a spacing benefit, but no difference between spaced or massed presentations was found in adults. However, this was due to the task being too simple for adults, so they reached ceiling performance for both conditions. The performance on the memory task was secondary in the study described above, as the primary research question was whether adults and children showed a bias in favour of the massed practice (which both adults and school-aged children did, despite consistently recalling more from spaced practice). Thus no further studies were conducted with appropriate difficulty levels for adults of the tasks, making the comparison less informative regarding age-related differences in spaced retrieval practice. Their findings do, however, indicate that adults and children benefit differently from spaced presentations of retrieval practice.

To our knowledge, no studies have directly compared learning performance in adults and children after practising retrieval in spaced or massed sessions (i.e., not spaced presentations like above). Instead, we can examine studies with similar methods to further explore potential developmental differences in the spacing effect between adults and children. For example, Ambridge et al. (2006) exposed children (3-4 years old) to sentences containing unfamiliar grammatical constructions in a single session or spaced across five consecutive days. By the end of the training period, children who practised the grammatical constructions across five days could recall significantly more correctly than those who practised in a single session. In contrast, adults learning unfamiliar English grammar in one session or sessions spaced across three days showed equal recall performance after training (Miles, 2014). Thus, in a language-related learning task, children showed an immediate benefit from spaced retrieval practice spanning several days, while adults did not. Unfortunately, Ambridge et al. (2006) did not include a long-term measure, meaning we cannot determine whether this initial developmental difference was maintained over a longer period.

Another approach to comparing the adult and child literature is to look at overall behaviour patterns in the different ages. In general, adults often show equal performance or a massed benefit immediately after the retrieval practice phase, regardless of whether a period of sleep occurs between retrieval occasions (e.g., Miles, 2014; Cepeda et al., 2009; Karpicke & Roediger, 2006), and a spacing benefit only emerges after a longer time-period have passed. In contrast, children more consistently show a spacing benefit immediately after the retrieval practice phase (Zigertman et al., 2015), but especially if a period of sleep occurs between retrieval occasions (e.g., Goossens et al., 2012; Moinzadeh et al., 2008), which then maintains in long-term measures. Based on these overall trends, it can be speculated that the developmental differences are greater if spaced retrieval practice sessions span across a sleep period. However, the lack of direct comparisons of adult and child populations using the same spacing methods and stimuli means firm conclusions cannot be drawn.

#### *1.3.2.2 Theoretical accounts of the developmental differences in spaced retrieval practice*

As with the general spacing literature in children, there is a distinct lack of accounts of spaced retrieval practice in development. We will now consider whether the theoretical accounts discussed in the context of adult research can be applied to data from children and can account for the tentative developmental differences outlined above.

The previously explored theoretical accounts of *contextual variability* and *retrieval effort* struggle to explain certain patterns of the child literature on spaced retrieval practice. For example, studies that report an early emerging spacing benefit in children do not indicate that retrieval was easier or that more overlapping contextual cues are available to aid initial retrieval after massed retrieval practice. Thus, according to these theoretical accounts, we would not expect a long-term benefit from spaced retrieval practice in children.

To apply the retrieval effort or contextual variability theories to the behaviours of children using spaced retrieval practice, we need to look further than performance at practice sessions. For instance, children have a smaller capacity to learn many words quickly (e.g. Cowan & Alloway, 2009; Schwarts & Terrell, 1983). Thus, one prediction could be that children might struggle to learn many words in massed sessions, making the retrieval attempts too hard and reducing the later benefit of retrieval practice. In contrast, if adults can learn more words in a single session (e.g. up to 40 word pairs; Bahrick & Hall, 2005; Cepeda et al., 2009), this may explain why adults practising novel words in a massed session can initially recall similar numbers of words as those in spaced sessions. Thus, the contextual variability and retrieval difficulty accounts could explain the performance pattern in spaced presentation studies; however, we will now examine the role of sleep (i.e., offline consolidation) and reconsolidation to explain long-term developmental differences in the spacing effect.

Children are especially sensitive to sleep periods occurring between retrieval attempts. For example, Schwarts and Terrell (1983) found that 12 young children (1 to 2 years of age) who practised novel words in spaced or massed sessions could learn more words if sessions were spaced out over several days. Specifically, words spaced in sessions over ten days were better recalled than those spaced out over five days. This finding would suggest that multiple consolidation opportunities from the intervening sleep led to improved recall than fewer nights of sleep. Later, Childers and Tomasello (2002) replicated these results using a larger sample size (36 children) and extended their findings by including a greater range of spacing schedules. Of particular interest is a finding that children learned and could recall more words if the spaced sessions spanned four consecutive days (one short session/day) than over two consecutive days (one longer session/day). This could be attributed to more consolidation and subsequent reconsolidation opportunities occurring over the four days compared to the two days, showing the importance of sleep in spaced retrieval practice in children.

This raises the question of whether spaced retrieval practice without intervening sleep periods is more beneficial than massed retrieval practice in children. Limited studies examine retrieval practice separated by more than intervening presentations but not crossing a period of sleep in children. One exception of a study that combined spaced sessions both within and across days is Seabrook, Brown and Solity (2005). Children in year 1 (mean age of five years and six months) practised reading ability (e.g. matching sounds with letters) every day over two weeks. They practised in three 2-minute sessions spread out throughout the school day (*spaced*) or one 6-minute session (*massed*). At the end of two weeks of within-day spaced or massed practice, the children with spaced practice had improved their reading ability significantly more than those with massed practice. This finding would suggest that spacing sessions throughout a single day are more beneficial for reading ability than massing sessions (even when combined with intervening overnight consolidation opportunities). However, an important caveat to this finding is that all children had a sleep opportunity between retrieval sessions, as the training schedule spanned sessions taking place every day for two weeks. Thus, while the shorter within-day spaced sessions enhanced reading abilities more than the longer sessions, all children benefited from overnight consolidation, meaning that there was no "pure" massed condition where all practice took place on a single day, without intervening sleep. Whether within-day spaced retrieval practice effectively enhances the long-term retention of novel words in children remains a question.

Interestingly, Childers and Tomasello (2002) also found that when the retrieval practice was separated into four shorter sessions, performance was the same if they were separated by 24 hours or three days apart. This would suggest that children benefit more from shorter sessions rather than longer ones. In contrast, adults can learn substantially more words in longer sessions without a period of consolidation to support further learning (Bahrick & Hall, 2005). A potential explanation for this could be prior knowledge allowing adults to relate novel words with words already embedded in the mental lexicon. Instead, children have less prior knowledge of words (due to having a smaller vocabulary), thus relying on offline consolidation processes to aid initial learning. James, Gaskell and Henderson (2019) directly tested whether adults and children benefit from prior knowledge and overnight sleep differently when learning novel words. While they found that both adults and children could use prior knowledge of words to aid recall before sleep, children showed a greater reduction in reliance on prior knowledge after a period of overnight sleep. Thus, this provides evidence that adults can rely on prior knowledge while children rely on offline consolidation to aid initial learning when it spans several days. Relating this to the spacing literature, it could mean that children need a period of offline consolidation between retrieval attempts as this would free capacity to continue to learn, resulting in higher initial performance compared to if no consolidation opportunities were given (e.g., in massed retrieval practice). Adults could instead use prior knowledge to effectively learn novel words, reducing the need for intervening sleep to aid initial learning from retrieval practice, resulting in an initial massed benefit.

Suppose sleep-based consolidation and reconsolidation play a role in the effects of spaced retrieval practice in both adults and children. In that case, this raises the question of whether the spacing effect's developmental differences are reduced if no intervening sleep occurs between retrieval occasions. Currently, there is no clear answer to this question as no within-day spaced retrieval practice studies directly compare adults and children using the same procedure and material. Thus, there is a significant empirical gap in the literature, resulting in potentially inappropriate methods being applied due to theoretical accounts not being tested thoroughly.

### 1.3.3 Summary of child literature

In the sections above, we have shown that children appear to rely on offline consolidation to integrate novel words with existing vocabulary, aiding word learning. This is a similar process as displayed in adults. However, and interestingly, children can show a greater benefit from sleep-based consolidation than adults, which has been demonstrated through behavioural evidence of greater improvements in word recall after a period of sleep (Weighall et al., 2016). Further, spaced retrieval practices have enhanced children's long-term retention of novel words. Interestingly, there appear to be subtle developmental differences in the spacing effect, where children can show an earlier emerging spacing effect than adults. However, the current evidence base for this is limited. Relatively few studies have looked at within-day spaced retrieval practice (i.e., without intervening sleep), particularly studies of word learning. Thus, it remains possible that differences in sleep-based consolidation could explain the different behaviour patterns in adults and children. A study comparing spaced versus massed retrieval practice without intervening sleep will allow us to explore if children benefit from the actual space between sessions or if sleep is needed for a benefit to emerge.

## 1.4 Concluding remarks and thesis overview

To conclude briefly before outlining the thesis, spaced retrieval practice has benefited long-term retrieval of novel words in adults and children. However, the findings in the literature and current theoretical accounts leave certain questions unanswered. In this thesis, I will address the questions outlined above. First, the effects of repeated retrieval practice with feedback can be similar to the observed spacing effect. Thus, to distinguish the two effects, removing feedback from within-day spaced retrieval attempts will allow us to focus on the spaces without the potential influence of additional learning opportunities from feedback. Second, a key issue with the existing spacing literature is the difficulty in distinguishing between sleep and spacing effects in both adult and child populations. Studies examining spaced retrieval practice often span sessions across several days, allowing offline consolidation to occur between retrieval attempts. Indeed, findings in the spacing literature can, in some cases, be explained by a sleep effect. Thus, removing the potential influence of sleep between spaced retrieval attempts would allow us to isolate the spacing effect. Further, by measuring retention one day after initial retrieval practice, we can determine how initial spaced or massed retrieval practice affects retention after offline consolidation and if adults and children show a difference in overnight change. Finally, by conducting studies using the same underlying methods and stimuli, we can provide a direct comparison of adult and child data, which does not currently exist.

Chapter 2 will focus on the research question: Does within-day spaced retrieval practice with and without feedback benefit retention of novel words before and after sleep and one week after initial exposure in adults? In three experiments, we examined how within-day spaced retrieval practice with or without feedback affected word learning performance (measured through cued recall, picture naming, and base animal matching tests) before (Day 1) and after (Day 2) a sleep period and one week (Day 7) after initial training in adults. Based on the within-day spaced retrieval practice literature outlined above, we expected to observe a spacing benefit before and after sleep and one week later, both with and without feedback in the retrieval attempts.

Chapter 3 will focus on the research question: Does within-day spaced retrieval practice (with feedback) practised in a classroom result in a spacing benefit in word learning before and after sleep and one week after initial exposure in children? We conducted two experiments where children learned and practised retrieval of novel animal names in two sessions spaced 3 hours apart or in a single session. In Experiment 4, children completed a follow-up session 24 hours later but not in Experiment 5. One week later, all children completed a final long-term retention test. Similar to our adult sample, we expected a spacing benefit to emerge on the first day and be maintained one day and one week later.

Chapter 4 will focus on the research question: Do adults and children show different benefits in word learning before and after sleep and one week after initial exposure? To close the empirical gap in the literature, we conducted two cross-experiment analyses to determine whether within-day spaced, and massed retrieval practice affects word learning differently for adults and children.

By attempting to answer these questions through five studies, we will provide evidence to help further close the empirical gap on the underlying processes of the spacing effect and whether within-day spaced retrieval practice can be an effective learning method to be used in educational settings.

# Chapter 2:

## Can within-day spaced retrieval practice enhance long-term recall of novel words in adults? Untangling the effects of feedback and time

The stimuli, data and analysis outputs for Experiments 1-3 are available at the Open Science Framework: https://osf.io/tvcm9/

The pre-registration for Experiment 3 is available at the Open Science Framework: https://osf.io/4cz67

## 2.1 Abstract

Spaced retrieval practice can effectively enhance the long-term retention of novel words in adults. However, the resulting spacing benefit can be similar to the effects of repeated retrieval practice with feedback and sleep's effects on word learning performance. In three experiments, we examined how within-day spaced retrieval practice with or without feedback affected word learning performance (measured through cued recall, picture naming, and base animal matching tests) before (Day 1) and after (Day 2) a sleep period and one week (Day 7) after initial training in adults. Without feedback (Experiment 1), there were minor improvements across retrieval sessions on the first day, but when feedback was provided on the retrieval occasions of the first day (Experiments 2 and 3), adults practising massed retrieval improved at a steeper rate than if practising in within-day spaced sessions. If a retrieval opportunity was given the day after learning, participants practising spaced retrieval caught up with the massed (Experiment 2), and participants maintained their performance one week later. Further, within-day spaced retrieval practice protected novel words from forgetting over the week later more than massed retrieval practice, but with performance nevertheless equivalent between conditions at the one-week test (Experiment 3). We discuss our findings in relation to existing theoretical accounts and educational implications. In short, the lack of clear spacing benefit in the three experiments indicates that the spacing effect is more complex than previously theorised, and aspects such as feedback and sleep deserve further examination.

## 2.2 Introduction

The ability to learn is crucial across the lifespan, and finding ways to optimise learning and memory retention has been of interest to psychologists for decades (e.g. Spitzer, 1939). Word learning is one critical process proposed to be supported by our memory systems (Davis & Gaskell, 2009). Understanding the mechanisms of word learning and revealing how to maximise the long-term retention of new words is key to informing effective educational practice. The present Chapter comprises three experiments examining whether the scheduling of post-learning retrieval practice can influence the retention of newly learned words when tested 24 hours and one week later.

### 2.2.1 Word learning

Accumulating evidence suggests that word learning is a lengthy procedure (e.g. Leach & Samuel, 2007) where new words undergo post-learning consolidation (i.e., the process by which fragile representations of words become strengthened in long-term memory and robust to interference; Rasch & Born, 2013). For example, the Complementary Learning Systems (CLS) framework (McClelland, McNaughton & O’Reilly, 1995; O'Reilly & Rudy, 2000; O'Reilly et al., 2014) proposes that, at a neurological level, consolidation is underpinned by a transfer of reliance from short-term hippocampal memory representation to long-term neocortical representation. Applying this framework to word learning, Davis & Gaskell (2009) proposed that novel words are initially encoded as events (i.e. episodic memory) mediated by hippocampal and medial-temporal lobe structures. Over time, and through consolidation, novel words are argued to become less hippocampally reliant and better integrated with lexical knowledge in the neocortex.

Consolidation of novel words is proposed through communication between hippocampal and neocortical regions via coordinated reactivation of memory traces during sleep (e.g. O'Reilly et al., 2014; O'Reilly & Rudy, 2000). More specifically, hippocampal memory traces are reactivated (represented as hippocampal ripples, 80-100Hz frequency) in a temporally synchronised manner with neocortical Slow Oscillations (at around 0.75Hz frequency; Timofeev et al., 2000), coordinated by sleep spindles (0.5-1.5 second bursts of 10-16Hz activity; Staresina et al., 2015), and together these electrophysiological events drive the consolidation process. Supporting this behaviourally, Dumay and Gaskell (2007) found that novel non-words become integrated with existing lexical knowledge (i.e., with lexical integration measured via a lexical competition effect between the new word and existing lexical competitors) after a period of sleep, but not after an equivalent period of wake. In addition, adults' performance on recall and recognition tasks improved more after sleep than wake, suggesting that sleep was crucial for improved implicit (lexical integration) and explicit knowledge of the words. Subsequent polysomnography studies have shown that sleep spindles and slow oscillation activity that occurs the night after learning new words can be associated with implicit and explicit measures of word learning the next day (e.g., Tamminen et al., 2010; Weighall et al., 2017). However, there is also evidence that learning new words influences subsequent sleep microstructure (e.g. Tamminen, Lambon-Ralph & Lewis, 2013; Gais et al., 2002), suggesting a bidirectional relationship between word learning and sleep. This raises the question of how different learning conditions may influence the sleep-associated consolidation process. Indeed, extensive literature suggests that the conditions of initial encoding can affect long-term memory. Such conditions include, but are not limited to, semantic properties

(Tamminen, Lambon-Ralph & Lewis, 2013), encoding strength (Pyc & Rawson, 2009), and, of particular interest here, *spaced retrieval practice.*

### 2.2.2 Spaced retrieval practice

Retrieval practice refers to the active recall of new information through testing, such as cued recall or multiple-choice questions. Numerous studies show benefits in long-term memory from active retrieval practice relative to passive restudy (for reviews, see Karpicke & Blunt, 2011; Roediger & Butler, 2011; and Adesope, Trevisan & Sundarajan, 2017; Rowland, 2014 for meta-analyses; although see Pickering, Henderson & Horner, 2021 for counterevidence). Retrieval practice is an effective learning aid in the laboratory and educationally relevant settings and has been increasingly recommended for use in classrooms (Surma, Vanhoyweghen, Camp & Kirschner, 2018). Of central importance here, retrieval practice has been claimed to be particularly effective when multiple retrieval attempts are "spaced" over time (i.e., separated by other learning items or over longer periods) relative to "massed" delivery (Latimier, Peyre & Ramus, 2021).

At this point, it is crucial to distinguish between "spaced learning" and "spaced testing". The former is where the whole learning phase is distributed over items/time points (and typically compared to "massed" learning in a single session). The initial learning is often slower in spaced over massed learning, as less learning takes place in each session (e.g. Cepeda et al., 2006; Shea et al., 2000). However, the rate of forgetting is slower than massed learning, resulting in better long-term retention of the learning material (even when the total time between the learning sessions and final retention test is controlled; Kornmeier, Sosic-Vasic & Joos, 2022; Cepeda et al., 2006). Spaced testing, on the other hand, refers to when learning is administered at a single point in time, and then a testing period is subsequently either spaced over items/time or massed at one point in time (typically immediately after learning). The overarching aim here is to address the question of whether learners benefit from spacing (as compared to "cramming") as a post-learning revision technique, with our testing protocol adhering to the principles of retrieval practice, henceforth referred to as "*spaced retrieval practice*".

A typical study exploring spaced retrieval practice consists of an initial exposure phase, followed by a retrieval practice phase (spaced or massed) which often involves corrective feedback, culminating in a final test/retention phase (often occurring after more extended time has passed, giving long-term memory measures). For example, in a seminal study, Bloom and Shuell (1981) explored the effects of spaced retrieval practice in word learning. First, they had participants learn unfamiliar French words in an initial learning session. Then, in the practice phase, participants returned and completed multiple-choice, fill-in and cued recall tasks to practice retrieving the novel words in short sessions on three consecutive days (spaced group) or in one long session on the same day (massed group). The final test phase consisted of an immediate recall test at the end of the practice phase and a long-term recall test 4 days later. At the immediate test, participants in both groups could recall similar levels of words. However, the spaced group outperformed the massed group four days later, suggesting that spaced retrieval practice led to a superior long-term recall. It should be noted that the distinction between spaced learning and spaced retrieval practice is not always straightforward. For example, of key importance to the present experiments, many spaced retrieval practice studies include corrective feedback following each retrieval attempt which provides an additional learning opportunity and makes it impossible to attribute any effects purely to spaced retrieval (e.g. Bell et al., 2014).

As well as having pedagogical relevance, examining the effects of spaced retrieval practice allows us to advance theories of the mechanisms that underlie long-term memory. As outlined in Chapter 1, *contextual variability theories* explain why spaced retrieval practice enhances long-term memory of novel words by arguing that the varied context of spaced retrieval practice provides a greater range of cues to aid later recall (Estes, 1955; Pashler et al., 2009; Maddox, 2016). Alternatively, according to the retrieval effort theories, more difficult retrieval in spaced sessions leads to better long-term retention (than easier retrieval in massed sessions) because of slower decaying memory traces (Pavlik & Anderson, 2008). These sets of theories predict an immediate benefit from massed retrieval practice (i.e., measured at the end of the practice period) as the similar context at each retrieval occasion would mean more overlapping cues aiding immediate retrieval, and because less time passed between retrieval attempts, retrieval would be easier.

However, there have been findings that do not accord with these theories. For example, in Bell et al. (2014), adult participants learned English-Swahili word pairs in an initial learning session. After initial exposure to the word pairs, participants practised them in a cued recall task until they could successfully recall 100% of the words (participants required 82 trials on average). Once they had reached the 100% correct recall criterion, the learning phase finished, and participants completed one retrieval practice session at varying schedules. It is important to note that the retrieval practice sessions included corrective feedback; thus, participants were provided with some additional learning in the retrieval practice phase. There were four schedules; massed (retrieval practice took place immediately after the learning phase finished), 12h spaced same-day (retrieval practice took place 12 hours after learning but on the same day, e.g. learning at 9am, retrieval practice at 9pm), 12h spaced overnight (retrieval practice took place 12 hours after learning but spanning a period of sleep, e.g. learning at 9pm, retrieval practice 9 am the day after), and 24h spaced (retrieval practice took place 24h after learning). In the retrieval practice session, the spaced groups that spanned a period of sleep (12h spaced overnight and 24h spaced) performed better than the other schedules, with the massed schedule remembering the fewest words. Additionally, the two 12h spaced groups differed in performance, with the 12h spaced overnight group remembering more words than the 12h spaced same-day group. This would indicate that the higher performance in the spaced groups with sleep could retrieve the words more easily and have more overlapping contextual cues to aid retrieval. According to the accounts outlined above, we would expect more forgetting in the better-performing groups after a longer time had passed. However, this is not what they found. Instead, the spaced groups that included a period of intervening sleep still outperformed the massed group 10 days later. Interestingly, the difference between the 12h spaced groups had disappeared. This suggests that we must consider retrieval difficulty, contextual variability, *and* sleep effects to explain the benefits of spaced retrieval practice.

Considering the role of sleep, Smith and Scarf (2017) proposed a *reconsolidation account* of the spacing retrieval practice benefit based on reviewing studies that incorporated spaces of 24 hours or more. They argue that the spacing of retrieval practice influences both the consolidation and reconsolidation of new material, such that partially consolidated memories undergo a reconsolidation process when retrieval practice is spaced across sleep. The claim is that a partially consolidated memory trace is reactivated during retrieval practice, placing it in a fragile state that elicits further consolidation during subsequent sleep, leading to memory strengthening (Smith & Scarf, 2017; Alberini, 2011). Thus, learning novel words through retrieval practice events that span periods of sleep may benefit from both consolidation and reconsolidation, resulting in more established memory traces and better long-term retention.

### 2.2.3 Separating the effects of spaces and sleep

Given the role of sleep in the consolidation and reconsolidation of new words and that many studies examining spaced retrieval practice include a period of sleep between tests, an important question is whether the spaced retrieval practice effect is observable without sleep. Studies that space retrieval practice blocks over a single day can go some way to addressing this question. However, surprisingly few studies have been conducted using this design.

Studies examining within-day spaced retrieval practice have often manipulated the number of intervening items within the retrieval session (as opposed to separating sessions across the day). For example, Karpicke and Roediger (2007) taught undergraduate students (18-22 years) word-pairs that were practised via retrieval using a cued recall task in massed (1 intervening item), equally spaced (5 intervening items), or expanding spaced (1, 5, then 9 intervening items) presentations. Ten minutes later, a cued recall task was administered. The results indicated that both spaced presentation schedules performed better than the massed schedule. Furthermore, the equally spaced condition produced the best performance two days later compared to the other conditions. Thus, these data suggest that a spaced retrieval practice effect can emerge when tests are separated by both very short (e.g. five intervening items) and longer (e.g. 12 hours in Bell et al., 2014) spaces. However, it is unclear whether the same mechanisms drive these effects. It is also difficult to draw comparisons between Karpicke and Roediger and Bell et al., given that the spaced retrieval practice condition in the former is akin to the massed condition in the latter (i.e., Bell et al. used random order presentation of their stimuli in all conditions, in the same way as Karpicke & Roediger did in their spaced condition).

Thus, studies that separate tests in time (rather than intervening items) would allow a cleaner comparison with studies that separate tests over periods of sleep.

More recently, Kroneisen and Kuepper-Tetzel (2021) explored whether a spacing benefit emerges when retrieval practice is spaced across more extended periods. In an initial training session, adults were exposed to Lithuanian-English word translations. Immediately after or after a 2-hour delay, participants practised retrieving the English word after being presented with the Lithuanian word as a cue. Participants were not provided feedback, so no additional encoding opportunities were provided after the initial training. Then, 12 hours later, a final test was delivered after a period of wake or sleep. In this test, participants who practised retrieving the words immediately after learning and underwent 12 hours of wake showed steeper forgetting than those who slept during the 12 hours. Interestingly, however, when the words were practised after a 2-hour delay, they did not show forgetting, regardless of whether the following 12-hour period included sleep or wake. The authors argued that delaying retrieval practice by 2 hours led to more robust memory representations than if the test was immediate. This suggests that spaced retrieval practice (without intervening sleep) is beneficial relative to massed retrieval practice and that subsequent sleep-associated consolidation effects can be attenuated by post-learning spaced retrieval practice. A possible explanation for this could be that retrieval practice acts as a form of consolidation (Antony et al., 2017). Specifically, due to the short time between retrieval attempts in the massed condition, retrieval could rely more on local short-term memory processes than coordinated communication between cortical regions (which is argued to be required for effective retrieval attempts, Antony et al., 2017). In other words, the retrieval in the massed condition was insufficient to elicit consolidation processes, resulting in forgetting if no sleep-based consolidation occurred immediately after. Instead, when practising retrieval after a 2-hour interval, the novel words underwent a consolidation process, resulting in less forgetting at later tests.

In sum, it remains unclear whether word learning benefits from spaced retrieval without sleep and whether such benefits maintain after sleep (following consolidation and reconsolidation). Furthermore, as noted, many studies (except Kroneisen & Kuepper-Tetzel, 2021; Karpicke & Roediger, 2007) examining spaced retrieval practice include feedback (i.e. providing additional learning opportunities; e.g. Kornmeier, Sosic-Vasic, Joos, 2022; Bell et al., 2014), reducing the extent to which they can inform theories that focus on the act of retrieval practice as central to the spaced retrieval practice effect. Tackling this is important if we are to better understand the mechanisms underlying word learning and realise how to optimally structure the revision of newly learned material (Surma, Vanhoyweghen, Camp & Kirschner, 2018).

### 2.2.4 The current experiments

Three experiments examined whether word learning can benefit from spaced retrieval practice without opportunities for consolidation/reconsolidation during sleep, manipulating the absence (Experiment 1) or presence (Experiment 2) of feedback in retrieval occasions, and the exclusion of an additional post-sleep retrieval practice opportunity (Experiment 3 was preregistered at https://osf.io/4cz67). All experiments taught participants the names (and associated photographic images) of real but rare animals (based on Fletcher et al., 2020). Natural material was selected to be more pedagogically meaningful and motivating for participants to learn than compared to novel nonsense words (as are often used in studies of word learning, e.g., Dumay & Gaskell, 2007). All experiments began with an initial exposure session, which was intentionally brief and intended to lay the foundation to observe change (i.e., improvements/maintenance/forgetting) over subsequent retrieval attempts. Immediately following this, three blocks of retrieval practice occurred. On Day 1, these blocks were either *massed* (all taking place back-to-back in a single session) or *spaced* (separated by 2 hours). By separating the spaced retrieval blocks by two hours, theoretically, retrieval should be more challenging (due to more time passing, allowing forgetting to occur; e.g., Pyc & Rawson, 2009) and/or the temporal context[1] should be more varied than if retrieval practice is massed (e.g. Karpicke, Lehman & Aue, 2014). Thus, if we observe a spaced retrieval benefit on Day 1, this cannot be attributed to sleep-associated consolidation or reconsolidation processes. We included two massed groups to counteract time-of-day effects: one group completed their exposure/testing in the morning, and the other in the afternoon. To determine whether performance after initial spaced or massed retrieval practice was affected after one night of sleep (e.g., whether a potential opportunity for consolidation/reconsolidation worked to enhance or reduce any benefit of spaced retrieval practice), Experiments 1 and 2 included a follow-up retrieval practice block on Day 2. Finally, to determine whether any benefit of spaced retrieval practice was held in the longer-term (importantly, if this is to be recommended as a tractable learning aid), all experiments included a final retrieval practice block one week after Day 1.

[1] The main variation in contextual cues in our experiments was assumed to be temporal context (i.e. more time passed between retrieval attempts, placing the participant in a different temporal context), as all blocks took place in the same, or different but identical, rooms. Both temporal (e.g. Karpicke et al., 2014) and environmental (e.g. Smith, Glenberg & Bjork, 1978; however see Imundo et al., 2020) variability can lead to benefits in later retention performance. However, an argument can be made that temporal context leads to greater benefits in later retention levels than environmental context during retrieval practice (e.g. Imundo et al., 2020; Karpicke, Lehman & Aue, 2014), supporting our reasoning for not manipulating the environmental context.

Extending Bell et al. (2014), we deployed three retrieval practice tests designed to activate phonological, orthographic and semantic aspects of word-form memory. We measured participants' memory through written responses in all tests, capturing subtle changes in how the reproduced memory traces matched the exposed form. We also manipulated the inclusion of feedback to determine whether additional learning opportunities are a necessary component for eliciting memory benefits in spaced retrieval practice. The tests comprised: (i) a cued recall task where participants were presented with the first two letters of a word and asked to provide the complete word, and (ii)  a picture-naming task where participants were presented with a picture and asked to provide the word semantically associated with it, and (iii) a base animal match task in which participants were asked to retrieve a newly learned word that was associated with a picture of a familiar animal (e.g., a picture of a hedgehog should elicit the written response "tenrec" - a tiny hedgehog with distinctive yellow and black stripes). The latter task was incorporated to encourage retrieval of the novel words in parallel with the retrieval of already known words with shared semantic properties. Previous studies have shown that learning novel material in a familiar context can aid learning (e.g. Lindsay & Gaskell, 2013). Therefore, this task was included to encourage and facilitate word learning (including making the new words relevant for future use) even without additional learning opportunities from feedback (in Experiment 1).

## 2.3 Experiment 1

Experiment 1 examined whether spaced retrieval practice (without further learning opportunities in the form of feedback at test) benefits memory for new words relative to massed retrieval practice. Crucially, we examined whether such benefits emerge before sleep and if they maintain 24 hours and one week later. Aligning with the retrieval difficulty and contextual variability theories, we predicted a benefit from spaced retrieval practice to emerge both by the end of Day 1 and in the following blocks on Days 2 and 7. However, if sleep is necessary to allow for reconsolidation (according to Smith & Scarf, 2017), we would not expect a benefit for spaced versus massed conditions in Experiment 1 at any time.

Feedback at the tests were removed here to isolate the role of spaced (versus massed) retrieval practice as distinguishable from spaced (versus massed) learning. This distinction is important as the underlying mechanisms are arguably different (see Smith & Kimball, 2010, who argue that feedback provides an additional learning opportunity and that the gap between initial learning/tests can affect the role of feedback). A difference in benefits from spaced retrieval practice with/without feedback can be found when comparing the findings of Bell et al. (2014) and Kroneisen and Kuepper-Tetzel (2021). Indeed, Bell et al. did not observe a benefit from spaced retrieval practice with feedback when separated 12 hours apart; however, Kroneisen and Kuepper-Tetzel reported a spacing benefit without feedback, regardless of sleep. Additionally, by ensuring that the space of retrieval practice is over separate sessions, we avoided comparing within-session and across-session spaces. Thus, here, we examine whether this same benefit is observed when tests are spaced out over the day (i.e., separated by 2 hours; in contrast to intervening items, as in Karpicke & Roediger, 2007, and a long gap of 12 hours, as in Bell et al., 2014) and whether this benefit remains after sleep and one week later.

Based on the immediate spacing benefit found in Karpicke and Roediger (2007) and Kroneisen and Kuepper-Tetzel (2021), we hypothesised that words learned in the spaced schedule would be better remembered (seen as higher performance in the three retrieval tasks) by the end of the retrieval practice blocks on Day 1, relative to words learned in the massed schedule. Second, we hypothesised that participants undergoing both schedules would improve their performance overnight (when tested on Day 2, relative to the final test on Day 1) as a consequence of both repeat testing and overnight sleep consolidation, but that the spacing benefit from Day 1 would maintain (e.g. Karpicke & Roediger, 2007). Finally, aligning with the contextual variability and retrieval effort hypotheses (i.e., Estes, 1955; Pashler et al., 2009; Maddox, 2016; Pyc & Rawson, 2009) we hypothesised that long-term memory performance (measured one week later) would be better for participants who practised retrieving words in the spaced schedule than in the massed schedule (assuming that the spacing benefits found by Karpicke & Roediger, 2007, and Kroenisen & Kuepper-Tezler, 2021, maintains after a more extended period has passed). The above hypotheses apply across all tests (i.e. cued recall, picture naming, base animal match).

### 2.3.1 Experiment 1 Methods

#### *2.3.1.1 Participants*

Sixty-three undergraduate students (mean age=19.68 years, SD=1.87) from the University of York, UK, were recruited. The sample size was determined based on a power analysis using G*Power 3 (Faul et al., 2007). To achieve an effect size of f=0.403 (based on the effect size and power, $\eta^2_p$=0.14, reported for a main effect of Group, i.e. massed, 12h spaced same day, 12h spaced overnight, 24h spaced, by Bell et al., 2014), we would need to recruit 18 participants for each retrieval schedule. We aimed for 21 participants/schedule to ensure potential dropouts/data loss would not mean a significant loss of power. Participants received cash or course credit as a reward for participation. Only native English speakers with no history of learning or reading disabilities were invited. Participants were randomly allocated to practice retrieval in one of three retrieval practice schedules when signing up for the study; Massed AM (n=21), Massed PM (n=21), or Spaced (n=21).

#### *2.3.1.2 Material*

The stimuli were names and pictures of real but rare animals intended to be novel to most participants. Nine rare animals were taken from Fletcher et al. (2020), with additional items added to produce a set of twenty-four stimuli items. The rare animals were paired to create two separate lists of 12 pairs. The lists were matched on syllabic length, and each animal pair was associated with a familiar "base" animal (e.g., e.g. a *tenrec* and an *echidna* both resemble a hedgehog; see Appendix A1 for the stimuli lists used in Experiments 1-3). In addition, new photographic images and audio recordings were produced for the complete set to ensure consistency across stimuli. Photographs were sourced from Google image searches; all photos depicted the animals in their natural settings, with the animals occupying at least 50% of the image. A female speaker with a neutral English accent produced the audio recordings of the animal names in a sound-proof booth. Participants were presented with the written word and photographic stimuli on computer screens using Microsoft PowerPoint, and they provided written responses in physical booklets provided by the experimenter. The stimuli were delivered in this manner since this experiment was also to be conducted with school-aged children in a classroom setting (i.e., Chapter 3).

#### *2.3.1.3 Design*

A mixed design was used: Schedule (Massed AM, Massed PM, Spaced) was a between-subjects independent variable, and Block (B1, B2, B3, B4, B5) was a within-subjects independent variable. The dependent variable for each task (cued recall, picture naming, and base animal match) was a matching percentage based on a Levenshtein distance between participants' responses and the correct answer (as described in Section 2.3.1.5).

#### *2.3.1.4 Procedure*

All participants completed one initial exposure session and five blocks of retrieval practice on three days. On the first day of the experiment (Day 1), participants started with an initial exposure session in the morning (Massed AM mean start time = 10:49, SD=00:36; Spaced mean start time = 11:36, SD=00:35) or in the afternoon (the Massed PM mean start time = 14:27, SD=00:43). The exposure session was followed by three massed or spaced blocks of retrieval practice (B1-3). In the massed schedules, B1-3 took place immediately after the exposure session. Participants were allowed to take short breaks between the blocks but were encouraged to keep them short. Participants in the spaced schedule started B1 immediately following the exposure session and were then dismissed, to return two hours later for B2 (B2 mean start time = 13:36, SD=00:35) and two hours after that for B3 (B3 mean start time = 15:36, SD=00:35). All participants returned 24 hours later (Day 2) for B4 (Massed AM Day 2 mean start time = 10:23, mean gap between B3 and B4 = 23 hours 33 minutes; Massed PM Day 2 mean start time = 14:23, mean gap between B3 and B4 = 23hours 56 minutes; Spaced Day 2 mean start time = 11:10, mean gap between B3 and B4 = 23hours 34 minutes). Finally, to measure long-term memory performance, all participants returned one week after Day 1 (Day 7) to complete B5 (Massed AM mean start time = 10:29, SD=00:37, Massed PM mean start time = 14:13, SD=00:51, Spaced mean start time = 11:30, SD=00:43). See Figure 3 for an illustration of the retrieval practice schedules. The exposure and retrieval practice blocks occurred in a controlled laboratory setting at the University of York, UK.

When participants arrived in the lab, they were informed of the experiment procedure and provided written consent for their participation. They were then given instructions for the exposure session and asked if they had any questions before commencing the first exposure task (prior knowledge task). Once the exposure session was complete, participants alerted the experimenter, who gave the first retrieval practice block instructions. Participants were told that the tests would be challenging and that they should try their best to recall the animal names when prompted. If they were unsure about their answer, they were instructed to guess and write down anything they could recall, even if it was only part of a name. It was important that an explicit and effortful retrieval attempt had occurred, as even brief retrieval attempts have been found to enhance learning and retention, regardless of whether the attempt is successful (Kornell & Vaughn, 2016; Kornell, Klein & Rawson, 2015; Barcroft, 2007). Once the first retrieval block was completed, participants again alerted the experimenter, who gave further instructions depending on their allocated schedule. Participants in the spaced schedule were instructed to leave the lab and return 2 hours later for the second retrieval block. Participants in the massed schedules were asked to complete the second retrieval block immediately but were allowed to take a brief break if needed (no more than 2 minutes). This was repeated for the third retrieval block, after which participants were dismissed and asked to return for the follow-up session the day after. On days 2 and 7, the retrieval blocks' procedure was identical, with the addition of payment being provided after the final block on Day 7.

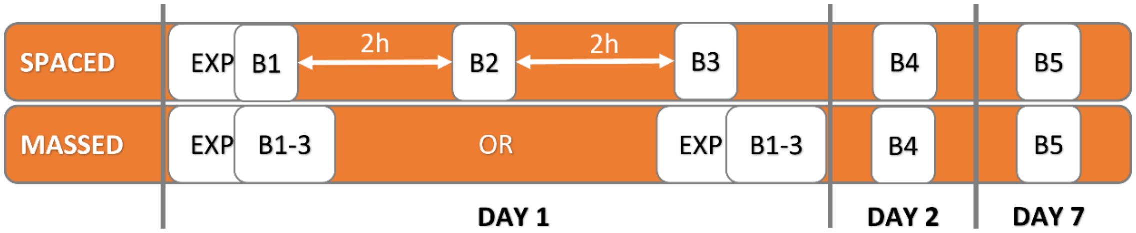


**Figure 3. Outline of the retrieval practice Schedules used in Experiments 1-3.** The top shows our Spaced schedule, where participants completed an exposure session followed by an immediate retrieval practice Block and two further blocks spaced 2 hours apart on the first day of the experiment. The bottom shows our Massed schedules, where participants completed an exposure session, immediately followed by three retrieval practice blocks back-to-back on the first day. The massed session took place in the morning (Massed AM) or afternoon (Massed PM) to counter potential time-of-day effects. Participants in all schedules then returned 24 hours later

(Day 2) and one week later (Day 7), allowing measures of overnight change in memory performance and long-term retention.

***Exposure session***

Three tasks (lasting no longer than 15 minutes) were designed to expose participants briefly to the names of the rare animals and their associated images. There were five auditory exposures of the animal name, four exposures to the written version of the animal name, four exposures to the picture, and only two complete exposures (i.e. picture and auditory and written name together).

*Prior knowledge task.* To determine whether participants were familiar with the animal names before the study, they were presented with an audio recording of each animal name and asked to indicate if they were familiar with it by circling YES or NO in their booklet. Each animal's name was presented once, in a random order, for 12 trials. If participants indicated prior knowledge of the animal, they were asked to write a brief definition (e.g. GIRAFFE = a horse-like animal that has a very long neck to reach tree tops) before moving on to the subsequent trial by pressing the space bar key. If they indicated prior knowledge and provided an accurate definition of the animal, the trials of the relevant animal were removed from analyses. No participants indicated prior knowledge of the animals, so no animal names were removed.

*Repetition task.* To allow participants to practice writing the animal name, they were presented with an audio recording and the written form of each animal name and asked to copy the name by writing it in their booklet. There was a time limit of 15 seconds for participants to copy the name, with the written name maintained on the screen, but they could move on faster by pressing the spacebar key. They were then presented with the name and accompanying picture for 3 seconds before the subsequent trial was initiated. Twelve trials occurred, one for each animal, in a randomised fixed order.

*Picture selection task (2AFC).* On each trial, participants were presented with two pictures of the newly learned animals, one target and one distractor, and the written version of the target animal's name (all presented simultaneously). The pictures were labelled A or B, and participants were asked to select which picture corresponded to the name by circling a letter in their booklet. The time limit for making their selection was 15 seconds before being presented with the correct answer for 3 seconds. The target and distractor animals were selected randomly. The trials were presented in a randomised fixed order.

***Retrieval practice blocks***

The retrieval practice blocks contained three tests always administered in the same order: cued recall, picture naming, and base animal matching. Each block took 10-15 minutes to complete. As initial exposure was limited to two complete exposures, the tests were expected to be challenging. Participants were encouraged to write down everything they could recall or guess if they were not confident. Unlike in the exposure session, the retrieval practice blocks provided no complete exposures to the stimuli to ensure no additional encoding opportunities would occur, ensuring that the experiment focused on spaced *retrieval practice* without the influence of additional learning opportunities.

*Cued recall*. This test measured participants' orthographic knowledge of the newly learned animal names. Participants were cued with the first two letters of each new animal name (e.g. TE) and asked to retrieve the full name and write it in their booklet (e.g. TENREC). Participants were given 20 seconds to write down their answers before the subsequent trial, but they could move on before the 20 seconds had passed by pressing the spacebar key if they finished writing their answers. The presentation order of the cues was a fixed random order, and each cue was presented once.

*Picture naming.* This test allowed us to measure the association between the picture and new orthographic forms. In a fixed randomised order, participants were presented with a picture of an animal and asked to retrieve and write down its name in their booklet. Participants were again given 20 seconds to write down their answers but could move on faster by pressing the spacebar key. Each picture was presented once for 12 trials in total.

*Base animal match*. This test was designed to encourage retrieval of the new animal name with a similar-looking already known animal to aid the integration of novel and existing information (inspired by Lindsay & Gaskell, 2013). Participants were presented with an already known animal (e.g. HEDGEHOG) and asked to retrieve and write down the name of a similar-looking animal they learned in the exposure session (e.g. TENREC). The task consisted of 12 trials, one for each new animal. After the base animal match task was finished, participants informed the experimenter, who gave instructions on the next step depending on which retrieval schedule the participant was assigned to and which block they had completed.

#### *2.3.1.5 Calculation of matching percentage (across-test dependent variable)*

Participants provided written responses in each retrieval practice test, allowing us to capture more subtle changes in spelling accuracy across tests (as opposed to scoring responses as correct/incorrect). Previous studies have shown that spelling accuracy is sensitive to encoding conditions (e.g. time between initial learning, recall tests, sleep; Kurdziel & Spencer, 2016).

An average matching percentage was calculated for each participant in the retrieval blocks. The matching percentage was based on the *Levenshtein Distance* (LD) between the target word and the participants' responses to determine the correctness of the participants' responses compared to the correct answer. LD determines the difference between two words by summing the number of deletions, insertions, and letter substitutions. The resulting LD number was then used to calculate a matching percentage (*matching percentage = 100 – ((LD)*100)/MaxLength),* which controls for the number of letters in the longer word of the two (i.e. MaxLength). Thus, the matching percentage indicates the correctness of the typed word, where 100% is an entirely correct recall, and 0% is no matching letters. For example, the LD between TENREC and TENGU is 3 (two substitutions: "g" for "r" and "u" for "e", one deletion: "c"), and the MaxLength is 6 (TENREC), resulting in a matching percentage of 50% (*matching percentage = 100 – (3*100)/6 = 50*). It is essential to mention that participants were always provided with the initial two letters of the animal name in the cued recall test. Thus, these would always be correct but would not accurately reflect retrieval performance. We removed the first two letters of the participants' responses and the correct answer before calculating the LD in the cued recall test to avoid affecting the results. In our earlier example, the matching percentage calculation would therefore be based on NREC (~~TE~~NREC) and NGU (~~TE~~NGU). While the LD has the same value of 3 (still two substitutions: "g" for "r" and "u" for "e", and one deletion: "c"), the MaxLength would be 4, resulting in a matching percentage of 25% *(matching percentage = 100 – (3*100)/4 = 25*), more accurately reflecting memory performance in this task.

#### *2.3.1.6 Statistical analyses*

To determine whether performance in the three retrieval tests on Day 1 differed for participants in the spaced and massed retrieval practice schedules, we calculated the average matching percentage for each participant as outlined before. The matching percentage was used as the dependent variable in a 3(Block: B1, B2, B3) x 3(Schedule: Massed AM, Massed PM, Spaced) Bayesian repeated-measures ANOVA. This analysis was run separately for the three tasks (cued recall, picture naming, and base animal matching). Further, additional analyses were conducted to determine whether overnight changes from Day 1 to Day 2 and long-term changes to Day 7 depended on the initial retrieval practice schedule. We ran a 3(Block: B3, B4, B5) x 3(Schedule: Massed AM, Massed PM, Spaced) Bayesian repeated-measures ANOVA for all tasks. B3 indicates performance at the end of Day 1, B4 indicates performance on Day 2, and B5 indicates performance one week later, on Day 7.

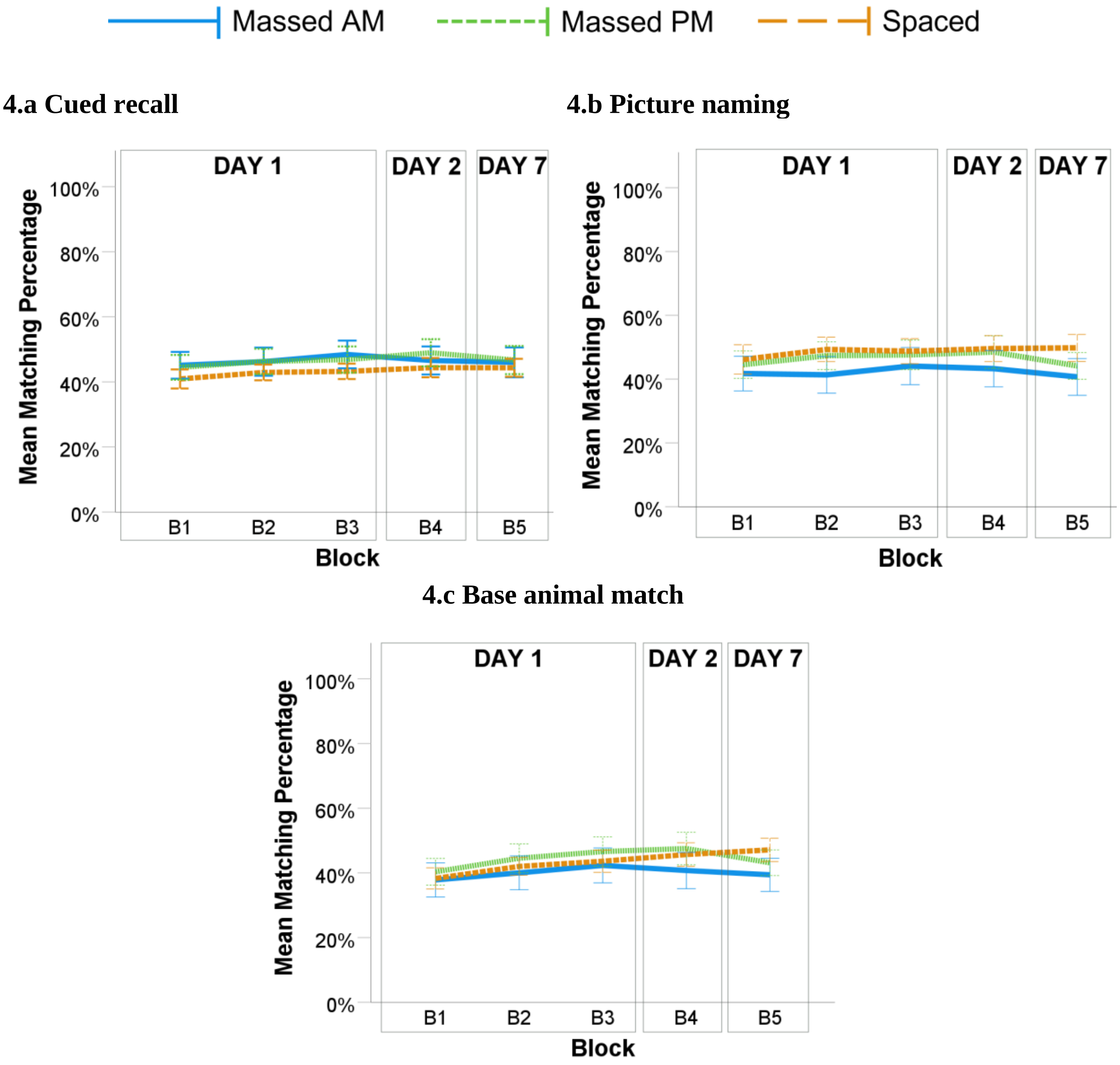
All analyses were run in JASP version 0.16 (JASP Team, 2022), an open-source software. As recommended by van Dorn et al. (2021) and Ly et al. (2016), prior distributions were not changed from the "default" distributions recommended for mixed ANOVA testing (see Rouder, Morey, Jordan & Province, 2012 for more information on the default priors). The alternative hypothesis was compared to the null hypothesis in all analyses, with the resulting *Bayes Factor* (BF) indicating support for either hypothesis in main effects ($BF_{10}$) or interactions ($BF_{incl}$). Per general guidelines on interpreting BF, a BF of 1 indicates no support for either hypothesis, BF between 1 and 3 are considered to be weak support, BF between 3 and 10 are considered moderate, and values above ten are considered to be strong support (van Dorn et al., 2021; Kass & Raftery, 1995; Jeffreys, 1998). In addition to reporting the relevant BF for effects and interactions, an error percentage (error%) will also be included, indicating the analysis's numerical robustness. Lower error percentages indicate greater stability in the numerical analysis; values below 20% are generally considered robust (van Dorn et al., 2021). Data and JASP outputs for all experiments in this thesis are available for open access at https://osf.io/tvcm9/.

### 2.3.2 Experiment 1 Results

Five of the 63 participants who signed up for the study did not return to the retrieval blocks per their schedule. Therefore, their data were excluded, resulting in 58 complete datasets used in the following analyses.

#### *2.3.2.1 The effect of within-day spaced (or massed) retrieval practice without feedback before sleep (B1-3)*

First, we analysed whether within-day spaced retrieval practice results in better word learning performance than massed retrieval practice before sleep. To do this, we ran a 3(Block 1-3)x3(Schedule) Bayesian repeated-measures ANOVA using the matching percentage (based on the LD calculation outlined in 2.3.1.5) in each word-learning test (cued recall, picture naming, base animal match). The mean matching percentage (± SD) for Experiments 1-3 can be seen in Table 1. The analyses revealed medium to strong support for changes in performance across blocks (i.e., main effect of Block) for the cued recall and base animal match tests (cued recall $BF_{10}$=4.2, error%=0.7; base animal match $BF_{10}$=4007.2, error%=0.84), but not in the picture naming test ($BF_{10}$=0.74, error%=0.57). Figure 4 shows improvements from B1 to B2 and B3 for all schedules in the retrieval tests. However, there was no support for the main effect of Schedule in any test (cued recall $BF_{10}$=0.54, error%=3.4; picture naming $BF_{10}$=0.57, error%=3.40; base animal match $BF_{10}$=0.48, error%=3.46). In addition, the analysis strongly supported the null hypothesis in Block*Schedule (cued recall $BF_{incl}$ = 0.06; picture naming $BF_{incl}$ = 0.09; base animal match $BF_{incl}$ = 0.05), indicating strong evidence against our hypothesis.


4.a Cued recall

4.b Picture naming

4.c Base animal match

**Figure 4. Experiment 1 results**. Graphs showing the mean matching percentage across blocks for the three groups in the cued recall (**4.a**), picture naming (**4.b**) and base animal match (**4.c**) tests in Experiment 1. Error bars represent ± 1SE.

### *2.3.2.2 The effect of within-day spaced retrieval practice without feedback on retention of novel animal names 24h and one week after initial exposure (B3-5)*

Our second set of analyses aimed to determine whether performance changed after 24 hours (including a period of sleep) and whether a long-term (one week) spacing benefit was present after initial within-day spaced retrieval practice without feedback. We conducted a 3(Block 3-5)x3(Schedule) Bayesian repeated-measures ANOVA using the matching percentage for the three retrieval practice tasks. The analysis revealed medium to strong support for the null hypothesis for Block (cued recall $BF_{10}$ = 0.13, error%=0.77; picture naming $BF_{10}$ = 0.23, error%=0.58; base animal match $BF_{10}$ = 0.12, error%=11.31), suggesting that participants did not change their performance across blocks on the three days (cued recall mean B3 = 46.2 (±15.1), B4 = 46.6 (±16.1), B5 = 45.7 (±16.3); picture naming mean B3 = 46.7 (±20.7), B4 = 47.2 (±21.5), B5 = 44.9 (±20.8); base animal match mean B3 = 44.2 (±19.2), B4 = 44.7 (±20.6), B5 = 43.2 (±18.4)). There was no support for the hypothesis for overall differences between Schedules in all tasks (cued recall $BF_{10}$ = 0.61, error%=3.79; picture naming $BF_{10}$ = 0.55, error%=3.31; base animal match $BF_{10}$ = 0.51, error%=3.36), suggesting that it is unlikely that participants performed differently depending on in which schedule they practised retrieval. Further, the analysis revealed no support for the hypothesis in the Block*Schedule interaction in cued recall and base animal match (cued recall $BF_{incl}$ = 0.54; base animal match $BF_{incl}$ = 1.21), but strong support for the null hypothesis in the picture naming test ($BF_{incl}$ = 0.12). Thus, there was no support for changes across blocks across the three days and no support for interaction, suggesting that performance was maintained at the same rate for all schedules, counter to our hypotheses.

**Table 1. Mean matching percentage (±SD) in the three retrieval practice tests in Experiments 1, 2 and 3.** The table contains the mean matching percentage in the retrieval blocks (B1-3 occurred on Day 1, B4 on Day 2, and B5 on Day 7) for participants in the different retrieval practice schedules (Massed AM/PM, Spaced) in Experiments 1, 2 and 3.

| | DAY 1 | | | DAY 2 | DAY 7 |
|---|---|---|---|---|---|
| | **B1** | **B2** | **B3** | **B4** | **B5** |
| **CUED RECALL** | | | | | |
| **Experiment 1** | | | | | |
| Massed AM | 45.1 (±16.4) | 46.2 (±17.4) | 48.4 (±17.0) | 46.6 (±17.3) | 46.0 (±18.4) |
| Massed PM | 44.5 (±16.6) | 46.3 (±17.0) | 46.9 (±17.5) | 48.9 (±18.3) | 46.7 (±18.9) |
| Spaced | 40.9 (±13.1) | 42.9 (±10.8) | 43.2 (±10.6) | 44.4 (±13.3) | 44.3 (±12.3) |
| **Experiment 2** | | | | | |
| Massed AM | 29.9 (±13.4) | 57.8 (±20.1) | 73.4(±19.7) | 78.3 (±18.6) | 77.3 (±17.4) |
| Massed PM | 26.9 (±12.6) | 46.5 (±20.5) | 64.7 (±25.0) | 67.3 (±26.7) | 67.9 (±26.6) |
| Spaced | 27.5 (±11.1) | 36.5 (±17.0) | 46.2 (±21.5) | 63.6 (±20.3) | 64.4 (±20.9) |
| **Experiment 3** | | | | | |
| Massed | 35.9 (±20.9) | 48.3 (±24.3) | 65.9 (±22.8) | | 51.1 (±23.8) |
| Spaced | 37.9 (±19.9) | 43.7 (±22.4) | 52.7 (±23.2) | | 51.5 (±24.9) |
| **PICTURE NAMING** | | | | | |
| **Experiment 1** | | | | | |
| Massed AM | 41.8 (±23.0) | 41.3 (±23.4) | 44.1 (±24.7) | 43.3 (±24.2) | 40.7 (±24.2) |
| Massed PM | 44.5 (±18.6) | 47.4 (±19.0) | 47.6 (±20.2) | 48.5 (±22.2) | 44.1 (±18.3) |
| Spaced | 46.2 (±24.4) | 49.3 (±17.5) | 48.3 (±17.9) | 49.7 (±18.7) | 49.8 (±19.4) |
| **Experiment 2** | | | | | |
| Massed AM | 29.9 (±13.4) | 57.8 (±20.1) | 73.4(±19.7) | 78.3 (±18.6) | 77.3 (±17.4) |
| Massed PM | 26.9 (±12.6) | 46.5 (±20.5) | 64.7 (±25.0) | 67.3 (±26.7) | 67.9 (±26.6) |
| Spaced | 27.5 (±11.1) | 36.5 (±17.0) | 46.2 (±21.5) | 63.6 (±20.3) | 64.4 (±20.9) |
| **Experiment 3** | | | | | |
| Massed | 40.0 (±21.3) | 53.2 (±22.6) | 69.5 (±21.1) | | 56.8 (±23.8) |
| Spaced | 40.9 (±18.8) | 49.1 (±21.9) | 59.1 (±22.5) | | 57.2 (±23.9) |
| **BASE ANIMAL MATCH** | | | | | |
| **Experiment 1** | | | | | |
| Massed AM | 37.8 (±22.4) | 40.0 (±22.2) | 42.3 (±22.9) | 40.8 (±23.8) | 39.4 (±21.8) |
| Massed PM | 40.4 (±17.8) | 44.6 (±19.5) | 46.6 (±19.9) | 47.5 (±21.9) | 43.1 (±17.3) |
| Spaced | 38.3 (±14.6) | 42.0 (±12.3) | 43.6 (±15.4) | 45.7 (±16.4) | 47.1 (±16.1) |
| **Experiment 2** | | | | | |
| Massed AM | 33.4 (±16.9) | 60.7 (±23.9) | 78.6 (±17.3) | 83.3 (±17.1) | 83.4 (±17.9) |
| Massed PM | 29.6 (±17.7) | 49.4 (±26.9) | 67.6 (±27.2) | 71.9 (±26.1) | 72.0 (±26.1) |
| Spaced | 27.7 (±13.6) | 42.9 (±16.8) | 54.6 (±18.6) | 71.4 (±16.6) | 69.1 (±17.1) |
| **Experiment 3** | | | | | |
| Massed | 33.5 (±20.2) | 49.8 (±24.4) | 65.1 (±24.1) | | 54.1 (±25.6) |
| Spaced | 33.8 (±18.6) | 43.1 (±21.5) | 53.9 (±22.8) | | 51.2 (±24.8) |

### 2.3.3 Experiment 1 Discussion

Experiment 1 explored whether within-day spaced retrieval practice would benefit word learning more than massed if no corrective feedback was provided. Regardless of the retrieval practice schedule (i.e., massed versus spaced), the matching percentage of the novel animal names recalled was remarkably similar, and this pattern was consistent across the three retrieval tests (cued recall, picture naming, base animal match). After the initial exposure session, participants performed at a level which would allow us to observe the potential effects of the key variables (i.e., far from both ceiling and floor, with the matching percentage of ~42% in the three tests). All participants showed small improvements across the retrieval practice tests on Day 1. However, in contrast to our hypotheses (and the findings of Karpicke & Roediger, 2007, and Kroneisen & Kuepper-Tetzel, 2021), participants who practised retrieval in the spaced schedule did not recall the animal names better than participants in the massed schedule. Furthermore, this pattern of results (i.e. no differences between spaced and massed schedules) was maintained on Day 2 and Day 7, indicating that no further improvements or forgetting occurred. These findings provide strong evidence against a spaced (relative to massed) retrieval practice benefit in the absence of feedback, questioning the contextual variability and retrieval effort theories.

The absence of a spaced retrieval practice benefit contrasts with findings from existing studies that also omitted feedback at tests (e.g., Karpicke & Roediger, 2007; Kroneisen & Kuepper-Tetzel, 2021). A key difference between the current study and previous studies is the level of initial exposure to the learning material. Kroneisen and Kuepper-Tetzel (2021) included two rounds of exposure as part of the initial training. Roediger and Karpicke (2007) only included one round but performance was still very high (98% correct recall in massed condition and 78% correct recall in spaced). The initially high performance brings an issue in itself as the results are susceptible to ceiling effects, potentially affecting the results by masking changes in performance. Additionally, Roediger and Karpicke (2007) reported a significant difference between spaced and massed retention levels, which is a limitation as the baseline level of performance was different, potentially affecting later outcomes. The current experiment purposely limited exposure to the stimuli (only 1 round of initial learning), resulting in lower levels of initial recall that are more amenable to observing subsequent condition differences (i.e. ~42% correct recall). Indeed, we did observe improved performance for both spaced and massed retrieval practice across blocks on day 1, which was not observed in the previous studies owing to their already high performance. Further, spaced retrieval practice without feedback has been argued to promote a spacing benefit by reducing forgetting; thus, initial performance may need to be high to observe such an effect.

Notably, Experiment 1 has implications for theories of spaced retrieval practice by indicating that we do not see a spacing benefit without feedback or intervening sleep between initial retrieval attempts. Thus, accounts such as contextual variability and retrieval difficulty should be reexamined to ascertain whether feedback/additional learning opportunities are required to elicit a spaced retrieval practice benefit. However, as mentioned, the current literature is not clear on this position. For example, studies often include intervening sleep *and* feedback, making it hard to distinguish whether feedback or sleep drives a spacing benefit. Therefore, in Experiment 2, we included feedback over within-day spaced (versus massed) retrieval practice opportunities to determine if a spacing benefit emerges before sleep and one week later.

## 2.4 Experiment 2

Including feedback following spaced retrieval practice attempts has been found to enhance later benefits in memory performance relative to both *massed retrieval practice* and *spaced restudy* (e.g. Agarwal et al., 2012; Carpenter, Pashler, Wixted & Vul, 2008; Karpicke & Roediger, 2007). Potentially, this means that spaced retrieval practice with feedback could be the ideal environment for promoting word learning and retention. However, the majority of studies exploring spaced retrieval practice with feedback involve intervening sleep, making it impossible to ascertain whether benefits arise from spaced retrieval practice or sleep-based consolidation and reconsolidation (e.g. Kim et al., 2019; Karpicke & Bauernschmidt, 2011; Logan & Balota, 2008; Bloom & Shuell, 1981). To address this limitation, Experiment 2 examined whether incorporating feedback in spaced retrieval attempts benefits word learning more than when retrieval attempts are massed when no intervening sleep occurs.

Currently, limited evidence suggests that a spacing benefit will emerge from within-day spaced retrieval practice. Karpicke and Roediger (2007) conducted a second experiment using the same design outlined earlier (i.e. comparing spaced and massed presentations of trials in a cued recall task) but included feedback (in the form of the correct answer) for all retrieval attempts. They again found better recall of the words practised in spaced retrieval presentations, but, crucially, the difference between the spaced and massed conditions was greater than if no feedback was provided. This suggests that spaced retrieval practice with feedback successfully encourages a spacing benefit in word learning. However, as mentioned, their design used spaced presentations (i.e., with intervening items) rather than separating sessions, making it difficult to draw direct predictions. Additionally, like in their first experiment (without feedback), the massed group performed significantly better than the spaced group immediately after the learning phase, potentially affecting later memory performance.

Also relevant to the current experiment are the findings of Bell et al. (2014). To reiterate, they found that within-day and across-sleep spaced retrieval practice resulted in an initial benefit when sleep (and feedback) was included, but this benefit disappeared in long-term (10 days) measures. This means that within-day spaced retrieval practice without sleep can have similar long-term benefits as spaced retrieval practice that includes sleep if feedback is included. However, as mentioned, the space between their sessions was very long (12 hours), and they used a paired-associate task instead of a word form learning measure. Additionally, participants practised the word pairs to a 100% correct recall criterion, potentially meaning that ceiling effects impacted later outcomes.

We included feedback in the form of the correct answer after each base animal match trial but not in the cued recall or picture naming tests. By only including feedback in the last test of each block, we aimed to reduce the risk of participants reaching ceiling levels (unlike in Karpicke & Roediger, 2007). Thus, feedback would be directly linked to retrieval attempts in the base animal match task but indirectly support retrieval in the cued recall and picture naming task. Based on previous studies that have reported benefits from spaced retrieval practice with feedback (e.g. Karpicke & Roediger, 2007; Bell et al., 2014), we hypothesised that words practised in spaced blocks would be better recalled than those practised in massed blocks by the end of Day 1 (i.e. B3). In addition, we hypothesised that the spacing benefit would maintain at the 24hr test (i.e. better performance from spaced schedule from B3 to B4), as seen in Karpicke and Roediger (2007). Finally, based on the long-term benefits of spaced retrieval practice with feedback from Bell et al. (2014), we expected participants in the spaced schedule to recall more words than those in massed schedules one week later (B5).

### 2.4.1 Experiment 2 Methods

#### *2.4.1.1 Participants*

Ninety undergraduate students (mean age=20.04 years, SD=1.94) from the University of York, UK, took part in this study. They were paid cash or received course credit for their participation. Only native English speakers with no history of learning or reading disabilities were invited. To more closely resemble studies that look at within-day spaced retrieval practice with feedback, we decided to increase the number of participants from Experiment 1 to 30 participants in each group (90 in total; based on Bell et al., 2014, as their design resembles ours).

#### *2.4.1.2 Design and material*

The design was identical to that of Experiment 1. The same stimuli were used as in Experiment 1 (i.e. names and pictures of rare animals) - however, participants performed at ceiling levels after piloting the 12 animal names with feedback. We were interested in whether feedback would elicit a spacing benefit while maintaining the other aspects of the design (e.g. lower levels of initial encoding). Therefore, we increased the number of animal names to 40, forming two lists with 20 names each. This resulted in slightly lower initial performance levels than in Experiment 1 but reduced the risk of ceiling effects. Further, all tasks were programmed in OpenSesame (Mathôt, Schreij & Schreij, 2012), allowing participants to provide their answers digitally and reducing potential transcription errors. In addition, we included the *Karolinska Sleepiness Scale* (KSS) to measure sleepiness and a *psychomotor vigilance task* (PVT) to measure attention (as described below).

#### *2.4.1.3 Procedure*

The procedure was identical to Experiment 1, with the following exceptions. First, all tasks were programmed in OpenSesame, such that participants provided their responses by typing them out on the computer (instead of in a physical booklet). Second, feedback was included in the base animal match task on Day 1. After providing their answer for each trial in the base animal match task, participants were presented with "CORRECT" in green or "INCORRECT" in red text. After 2 seconds, the text disappeared, and the correct answer was presented for 3 seconds before the onset of the subsequent trial. To ensure all participants received equal levels of exposure, they were presented with feedback regardless of whether their answers were correct or incorrect. Third, the PVT task was included to ensure differences in levels of attention could not account for any observed differences in the scheduling conditions. We used Reifman et al.'s (2018) task but shortened the duration from 10 to 5 minutes. The task was completed at the end of the final retrieval block of Day 1, Day 2 and Day 7. Participants were given instructions on how to complete the task, where a timer was presented on the screen at random intervals from 2'000ms to 10'000ms, and they had to press a mouse button as fast as possible to stop the timer from counting up. Reaction time (RT) and lapses (responses exceeding 500ms) were used in the analyses. Finally, we also administered the KSS (Miley, Kecklund & Åkerstedt, 2016; Shahid et al., 2011; Åkerstedt & Gillberg, 1990) at the start of all five retrieval blocks to determine whether varying levels of sleepiness affected retrieval performance. Participants were presented with a scale of 1-9 (*1 = Extremely sleepy, almost falling asleep; 9 = Extremely alert*) and asked to indicate how they felt 5 minutes before.

#### *2.4.1.4 Statistical analyses*

Data was prepared as in Experiment 1, with matching percentages calculated as previously described. Similarly, all core analyses were identical, and outliers were removed accordingly.

For the PVT analyses, we calculated the number of lapses (RT > 500ms) for each participant on days 1, 2, and 7. We also calculated the average reaction time for each participant on each day. Using these data, we conducted two 3(Day: Day 1, Day 2, Day 7) x 3(Schedule: Massed AM, Massed PM, Spaced) Bayesian repeated-measures ANOVAs, using the number of lapses and average reaction time as dependent variables in the separate analyses.

To compare sleepiness ratings between the groups in the different sessions, we used the KSS rating from each participant in a 5(Block: B 1, B 2, B 3, B 4, B 5)x3(Schedule: Massed AM, Massed PM, Spaced) Bayesian repeated-measures ANOVA.

### 2.4.2 Experiment 2 Results

Four participants (2 from Massed AM and 2 from Massed PM) were identified as outliers in the cued recall task. In addition, one participant from Massed PM was an outlier in the picture naming task. Finally, two participants were identified as outliers (1 from Massed AM and 1 from Massed PM) in the base animal match task. Therefore, the data from the identified participants were removed from the relevant analyses.

#### *2.4.2.1 The effect of within-day spaced (or massed) retrieval practice with feedback before sleep (B1-3)*

We analysed whether within-day spaced retrieval practice results in better word learning performance than massed retrieval practice before sleep. To do this, we ran a 3(Block 1-3)x3(Schedule) Bayesian repeated-measures ANOVA using the matching percentage in each word-learning task (cued recall, picture naming, base animal match). The analysis revealed overwhelming support for a main effect of Block in all tasks (cued recall $BF_{10}$=1.3e+41, error%=0.88; picture naming $BF_{10}$=4.3e+57, error%=1.1; base animal match $BF_{10}$=1.1e+52, error%=0.62). There was also strong support for a main effect of Schedule (cued recall $BF_{10}$=30.5, error%=0.88; picture naming $BF_{10}$=7.42, error%=1.15; base animal match $BF_{10}$=5.31, error%=0.72). Posthoc tests were conducted to explore this effect further. The results of the posthoc analyses can be seen in Table 2, and they suggest that Massed AM and PM performed higher than Spaced, but Massed did not differ depending on if they did their session in the morning or afternoon. Posthoc analyses of the Block variable reveal overwhelming support for a change in performance between all blocks, which, as seen in Figure 5, is due to overall improvements in performance.

**Table 2. Posthoc analyses for Schedule and Block variables on Day 1 (B1-3).** The posthoc analyses provide an unadjusted $BF_{10}$; however, the posterior odds (PO) correct for multiple comparisons by setting the prior odds (i.e. the probability of no effect) to 0.5 (Bergh et al., 2019; Westfall, Johnson & Utts, 1997).

| **SCHEDULE VARIABLE** | | **Unadjusted $BF_{10}$** | **Posterior Odds (PO)** | **Error %** |
|---|---|---|---|---|
| **Cued recall** | | | | |
| Massed AM | vs Massed PM | 0.99 | 0.58 | 2.9e-6 |
| | vs Spaced | 13358.7 | 7846.9 | 1.3e-11 |
| Massed PM | vs Spaced | 5.90 | 3.46 | 5.6e-7 |
| **Picture naming** | | | | |
| Massed AM | vs Massed PM | 2.25 | 1.44 | 1.6e-6 |
| | vs Spaced | 3561.9 | 2092.2 | 3.9e-10 |
| Massed PM | vs Spaced | 0.59 | 0.35 | 5.5e-6 |
| **Base animal match** | | | | |
| Massed AM | vs Massed PM | 1.16 | 0.68 | 3.0e-6 |
| | vs Spaced | 1205.7 | 708.3 | 1.0e-9 |
| Massed PM | vs Spaced | 0.91 | 0.54 | 3.8e-6 |

| **BLOCK VARIABLE** | | | | |
|---|---|---|---|---|
| | | **Unadjusted $BF_{10}$** | **Posterior Odds (PO)** | **Error %** |
| **Cued recall** | | | | |
| B1 | vs B2 | 4.1e+16 | 2.4e+16 | 1.0e-22 |
| | vs B3 | 5.2e+23 | 3.1e+23 | 1.3e-27 |
| B2 | vs B3 | 1.4e+16 | 7.9e+15 | 3.2e-22 |
| **Picture naming** | | | | |
| B1 | vs B2 | 9.7e+22 | 5.7e+22 | 4.7e-31 |
| | vs B3 | 1.3e+33 | 7.4e+32 | 4.3e-38 |
| B2 | vs B3 | 5.4e+23 | 3.2e+23 | 6.5e-31 |
| **Base animal match** | | | | |
| B1 | vs B2 | 3.9e+22 | 2.3e+22 | 1.8e-30 |
| | vs B3 | 4.9e+31 | 2.9e+31 | 6.8e-37 |
| B2 | vs B3 | 1.3e+16 | 7.5e+15 | 3.0e-22 |

To determine whether performance across the blocks was affected by the retrieval schedule, we focused on the Block*Schedule interaction. There was overwhelming support for the Block*Schedule interaction in all three tests (cued recall $BF_{incl}$=2.7e+7; picture naming $BF_{incl}$=607682.9; base animal match $BF_{incl}$=1537.0). Figure 5 illustrates this: while all participants improved from Block 1 to 2 and 3, those in the Massed AM and PM schedules improved at a steeper rate than those in Spaced, counter to the hypotheses.

#### *2.4.2.2 The effect of within-day spaced retrieval practice with feedback on retention of novel animal names 24h and one week after initial exposure (B3-5)*

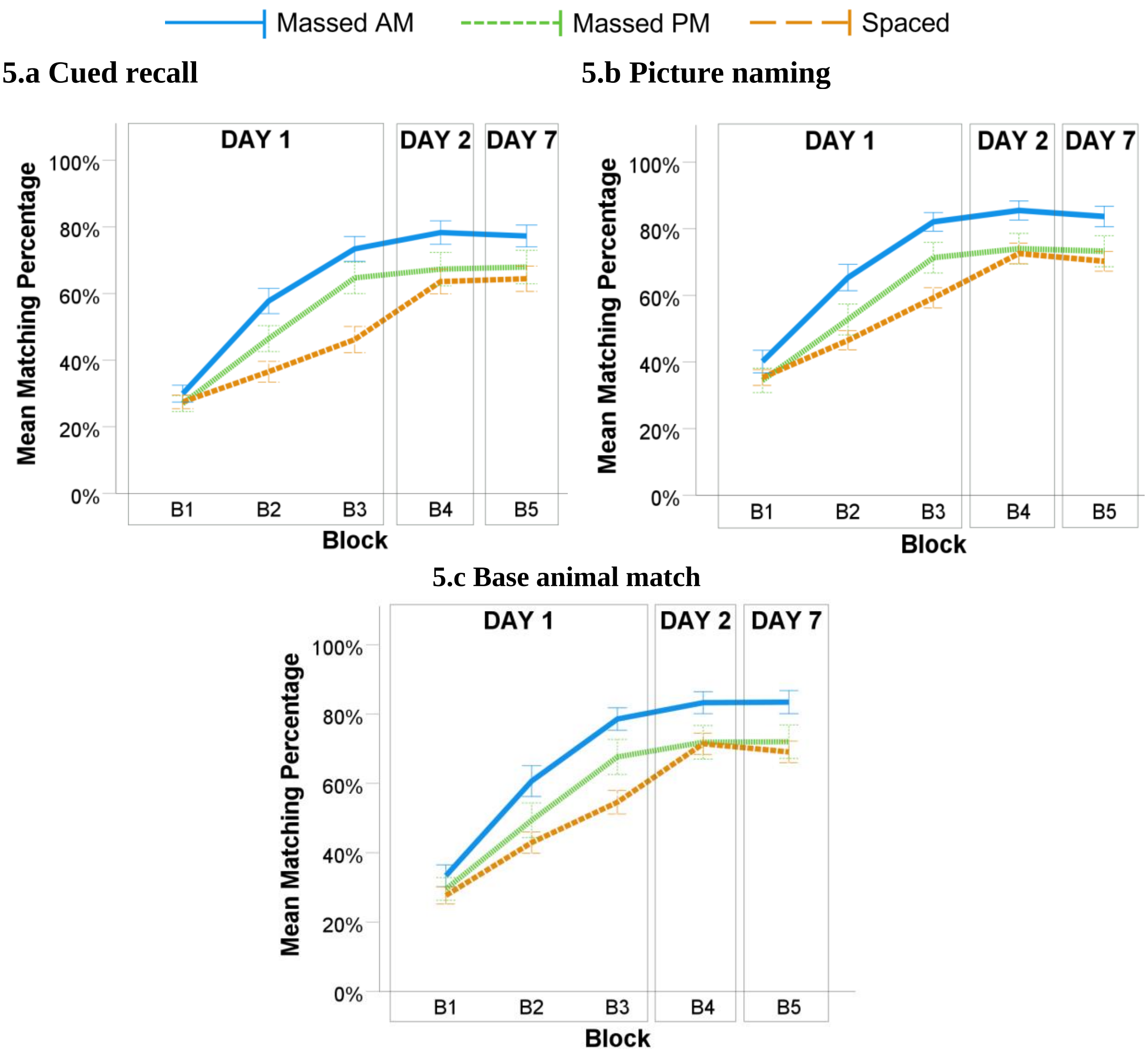
Our second set of analyses aimed to determine whether performance changed after one night of sleep and/or one week later (with feedback). We ran a 3(Block 3-5)x3(Schedule) Bayesian repeated-measures ANOVA using the matching percentage for the three retrieval practice tests. The analysis revealed overwhelming support for a main effect of Block (cued recall $BF_{10}$=1.8e+10, error%=0.71; picture naming $BF_{10}$=1.2e+12, error%=0.51; base animal match $BF_{10}$=1.2e+18, error%=0.7). There was also moderate support for an effect of Schedule (cued recall $BF_{10}$=8.50, error%=24.98; picture naming $BF_{10}$=9.59, error%=30.22; base animal match $BF_{10}$=5.61, error%=11.88). The posthoc analyses revealed overwhelming support for a change in performance from B3 to B4 and B5, indicating that performance improved from Day 1 to 7. However, there was support for the null hypothesis from B4 to B5 (i.e., Day 2 to Day 7) in two out of three tasks. Table 3 shows that these analyses support an overnight improvement from Day 1 to Day 2 and maintained performance from Day 2 to Day 7 (except for picture naming, where some forgetting occurred). Posthoc analyses for Schedule indicated that the Massed AM condition performed higher than the Spaced condition in all tests and that the Massed PM condition performed higher than Spaced for cued recall. While we expected a difference between the schedules, the results contrasted our hypothesis of a spacing benefit (because Massed performed higher than Spaced consistently).

**Table 3. Posthoc analyses for Schedule and Block variable on Days 1-7 (B3-5).** The posthoc analyses provide an unadjusted $BF_{10}$; however, the posterior odds (PO) correct for multiple comparisons by setting the prior odds (i.e. the probability of no effect) to 0.5 (van den Bergh et al., 2020; Westfall, Johnson & Utts, 1997).

| **SCHEDULE VARIABLE** | | | | |
|---|---|---|---|---|
| | | **Unadjusted $BF_{10}$** | **Posterior Odds (PO)** | **Error %** |
| **Cued recall** | | | | |
| Massed AM | vs Massed PM | 5.8 | 3.4 | 4.0e-7 |
| | vs Spaced | 446568.9 | 262315.0 | 2.0e-13 |
| Massed PM | vs Spaced | 2.1 | 1.2 | 1.7e-6 |
| **Picture naming** | | | | |
| Massed AM | vs Massed PM | 43.45 | 25.52 | 6.7e-8 |
| | vs Spaced | 1.8e+7 | 1.1e+7 | 3.4e-15 |
| Massed PM | vs Spaced | 0.64 | 0.38 | 5.1e-6 |
| **Base animal match** | | | | |
| Massed AM | vs Massed PM | 25.43 | 14.94 | 1.0e-7 |
| | vs Spaced | 1.8e+6 | 1.0e+6 | 1.9e-14 |
| Massed PM | vs Spaced | 0.54 | 0.54 | 6.0e-6 |

| **BLOCK VARIABLE** | | | | |
|---|---|---|---|---|
| | | **Unadjusted BF$_{10}$** | **Posterior Odds (PO)** | **Error %** |
| **Cued recall** | | | | |
| B3 | vs B4 | 6.3e+6 | 3.7e+6 | 1.6e-12 |
| | vs B5 | 371429.2 | 218177.9 | 2.5e-11 |
| B4 | vs B5 | 0.12 | 0.07 | 8.7e-5 |
| **Picture naming** | | | | |
| B3 | vs B4 | 4.4e+9 | 2.6e+9 | 1.7e-15 |
| | vs B5 | 21594.8 | 12684.8 | 1.6e-10 |
| B4 | vs B5 | 12.34 | 7.25 | 5.7e-7 |
| **Base animal match** | | | | |
| B3 | vs B4 | 7.2e+11 | 4.2e+11 | 1.9e-18 |
| | vs B5 | 3.8e+8 | 2.2e+8 | 2.3e-14 |
| B4 | vs B5 | 0.28 | 0.16 | 4.2e-5 |

To determine whether the changes across Blocks depended on which schedule participants were in, we focused on the Block*Schedule interaction. The analysis revealed overwhelming support that the changes across Block depend on the retrieval Schedule across all tests (cued recall BF$_{incl}$=1.1e+7; picture naming BF$_{incl}$=8.3e+8; base animal match BF$_{incl}$=2.7e+9). Figure 5 illustrates that from B3 to B4, the improvement in performance for the Spaced condition (cued recall mean change = 17.4 (±10.4), picture naming mean change = 13.3 (±6.3), base animal match mean change = 16.9 (±6.3)) was greater than for the Massed AM (cued recall mean change = 4.9 (±9.2), picture naming mean change = 3.4 (±5.1), base animal match mean change = 4.7 (±7.3)) and Massed PM conditions (cued recall mean change = 2.6 (±9.7), picture naming mean change = 2.7 (±5.7), base animal match mean change = 4.2 (±5.9)), with the massed groups instead maintaining their performance.



**5.a Cued recall**

**5.b Picture naming**

**5.c Base animal match**

**Figure 5. Experiment 2 results.** Graphs showing the mean matching percentage for participants in the three schedules across blocks on Days 1-7 in the cued recall (**5.a**), picture naming (**5.b**) and base animal match (**5.c**) tests. Error bars represent ± 1SE.

### *2.4.2.3 PVT and KSS*

Six participants had to be removed from the PVT analyses as technical problems with the software led to data not being saved correctly on at least one occasion, and two participants were removed due to being outliers, leaving 62 datasets. The analysis of the number of lapses (<500ms) revealed strong support for the null hypothesis in all variables (Block $BF_{10}$=0.10, error%=1.35; Schedule $BF_{10}$=0.11, error%=0.67; Block*Schedule $BF_{incl}$=0.21). The analysis of reaction time revealed strong support for changes across Blocks ($BF_{10}$=230.1, error%=3.09) but no support for differences between Schedules ($BF_{10}$=1.48, error%=1.96) and support for the null hypothesis in the Block*Schedule interaction ($BF_{incl}$=0.25). On average, participants showed slower reaction times across blocks, but this did not depend on which schedule they were assigned to, suggesting that the changes in attention did not account for the group differences in the other tasks.

The analysis of sleepiness ratings (KSS) revealed strong support for the null hypothesis in all measures, indicating that sleepiness ratings did not differ across Blocks ($BF_{10}$=0.04, error%=0.50), Schedule ($BF_{10}$=0.39, error%=2.16) and that changes in sleepiness ratings did not differ across Blocks depending on retrieval Schedule (Block*Schedule $BF_{incl}$=0.03). Thus, varying levels of attention or sleepiness did most likely not affect performance in the other tasks.

### 2.4.3 Experiment 2 Discussion

Experiment 2 examined whether within-day spaced retrieval practice (with feedback) benefits newly learned word recall compared to massed retrieval practice. The performance pattern differed dramatically from Experiment 1. All schedules showed clear improvements over the blocks on Day 1, suggesting that feedback is essential in facilitating word learning after initially limited exposure. However, contrary to our first hypothesis, participants in the spaced retrieval practice schedule did not perform better than those in massed schedules on Day 1. Instead, participants in the Massed schedules showed steeper improvements across B1-3 than those in the Spaced schedule. Furthermore, whilst participants in the Spaced Schedule showed bigger improvements from Day 1 to Day 2 than the Massed Schedules, this benefit only had the effect of helping the spaced group reduce their Day 1 disadvantage. We, therefore, also reject our predicted spacing benefit on Day 2. Finally, all participants maintained performance between Day 2 and Day 7, again indicating no spaced retrieval practice benefit after one week, thus rejecting our final hypothesis. In short, these findings provide further evidence against a spaced retrieval practice benefit prior to sleep (and after sleep), even when additional encoding opportunities through feedback were included.

Whilst the larger overnight improvements in the Spaced schedule could be a consequence of lower initial performance levels leaving more room for overnight improvement (see Section 2.6.1 for further discussion of this), another explanation could relate to the role of reconsolidation. Specifically, the steeper rate of improvement from Day 1 to Day 2 for the Spaced condition points toward a more nuanced spacing benefit and aligns with the view that spaced retrieval practice (with feedback) benefits more from overnight consolidation and subsequent reconsolidation than massed retrieval practice (Smith & Scarf, 2017). Sleep-based consolidation between Days 1 and 2 may have partially consolidated the new animal names along with the feedback (Tamminen et al., 2012). Then, in the subsequent retrieval block on Day 2, reconsolidation processes may have strengthened the association between feedback and animal name, resulting in the overnight improvements observed (Smith & Scarf, 2017).

However, it is difficult to reconcile this explanation with the absence of a spacing benefit on Day 7. According to the contextual variability and retrieval difficulty theories, we expected participants in the spaced schedule to remember more words, especially in the measures taking place one week after initial retrieval practice. Including the retrieval practice block on Day 2 could have elicited reconsolidation in both conditions (effectively providing a spaced retrieval practice opportunity even in the massed condition), therefore reducing our chances of observing a spaced retrieval practice benefit on Day 7. We address this directly in

Experiment 3, allowing us to isolate potential longer-term effects of within-day spaced retrieval practice.

## 2.5 Experiment 3

As the post-sleep retrieval opportunity on Day 2 could have masked benefits from the spaced retrieval practice schedule on Day 1, we removed B4 to determine whether within-day spaced retrieval practice leads to long-term benefits one week later without the potential influence of reconsolidation processes in post-sleep retrieval blocks. Thus, the blocks in the final design were B1-3 on Day 1, and B5 on Day 7. The methods and hypotheses were preregistered at https://osf.io/4cz67[2].

Beyond the omission of the Day 2 Block, there were two key methodological departures from Experiment 2. First, due to the Covid-19 pandemic, we were required to adapt the procedure to adhere to government guidelines. As such, we ran the experiment online, using Gorilla.sc. Second, since participants practising retrieval in the Massed AM and Massed PM schedule showed the same pattern of change across blocks, we recruited a single Massed schedule group (half of which did the initial training block in the morning and the other half in the afternoon).

Three hypotheses were pre-registered based on previous literature, the findings of Experiment 2, and the potential influence of reconsolidation (Smith & Scarf, 2017):

1) On Day 1, there will be improvements in all measures across retrieval practice blocks (i.e. as reflected by a main effect of Block).

[2] Note that the pre-registration document contains the term “Test” (replaced with “Block” here) and “Group” (replaced with “Schedule” here).

2) On Day 1, there will be an interaction between Block and Schedule for all measures, with participants in the Massed schedule improving at a steeper rate than in the Spaced schedule.
3) When comparing the final performance on Day 1 (B3) and long-term performance one week later on Day 7 (B5), if within-day spaced retrieval practice is more effective in promoting long-term memory than massed retrieval practice, we expect a Block*Schedule interaction in all measures (i.e., participants in the Spaced schedule will show greater improvements or less forgetting from B3 to B5 than participants in the Massed schedule).

### 2.5.1 Experiment 3 Methods

#### *2.5.1.1 Participants*

Sixty undergraduate students (mean age = 22.68, SD=4.73) were recruited from an online recruitment platform used at the University of York (SONA). They were paid compulsory research hours or an Amazon voucher for taking part. Half were assigned to practice retrieval in the Spaced schedule and the remaining half in the Massed. Half of the Massed retrieval schedule participants took part in the morning and the remaining half in the afternoon.

Due to the online nature of the study, we continuously monitored participants' attention by including *attention checks* (described in Section 2.5.1.4) to ensure full attention was given to the tasks. In the case of a missed attention check, the participant was prevented from continuing with the study. Data collection continued until the desired sample size of 60 participants had been met.

#### *2.5.1.2 Material*

The material was the same as Experiment 2 (i.e. names and pictures of 40 rare animal names in 2 lists). Instead of presenting the tasks in OpenSesame, they were programmed using the Gorilla Experiment Builder (https://gorilla.sc/) to allow online data collection; however, the task presentation was identical. In addition, due to technical restrictions with online testing, we substituted the PVT task with a sustained attention task (SART).

#### *2.5.1.3 Design*

The design was the same as Experiments 1 and 2, except for the Schedule variable only containing two categories (Massed and Spaced) and the Block variable only containing four categories (B1, B2, B3, B5).

#### *2.5.1.4 Procedure*

Except for running online, the procedure was the same as Experiment 2. To ensure participants completed the sessions on their given schedule, they were sent an email with the schedule and regular email reminders. They were also informed that if they did not start their sessions within an appropriate time window, they could not continue their participation and would not be paid.

In SART, participants were presented with a single letter on the screen for 500ms and instructed to quickly press the spacebar key, except when an "X" appeared. When an "X" appeared, they should not press any button but wait until the subsequent trial started. Ten rounds of 8 target letters and 2 Xs appeared at an interstimulus interval of 900ms for 100 total trials.

Due to online testing, there was an issue of ensuring participants paid attention to the tasks, which we countered by including attention checks. A cartoon picture of an object (e.g. table) and the text "*To ensure that you are paying attention in this task, please type in the* *object's name below" were* presented once in the picture naming task in each retrieval block. The objects were easy to identify (apple, fork, table, tree), and it was assumed that all participants could name them easily.

#### *2.5.1.5 Statistical analyses*

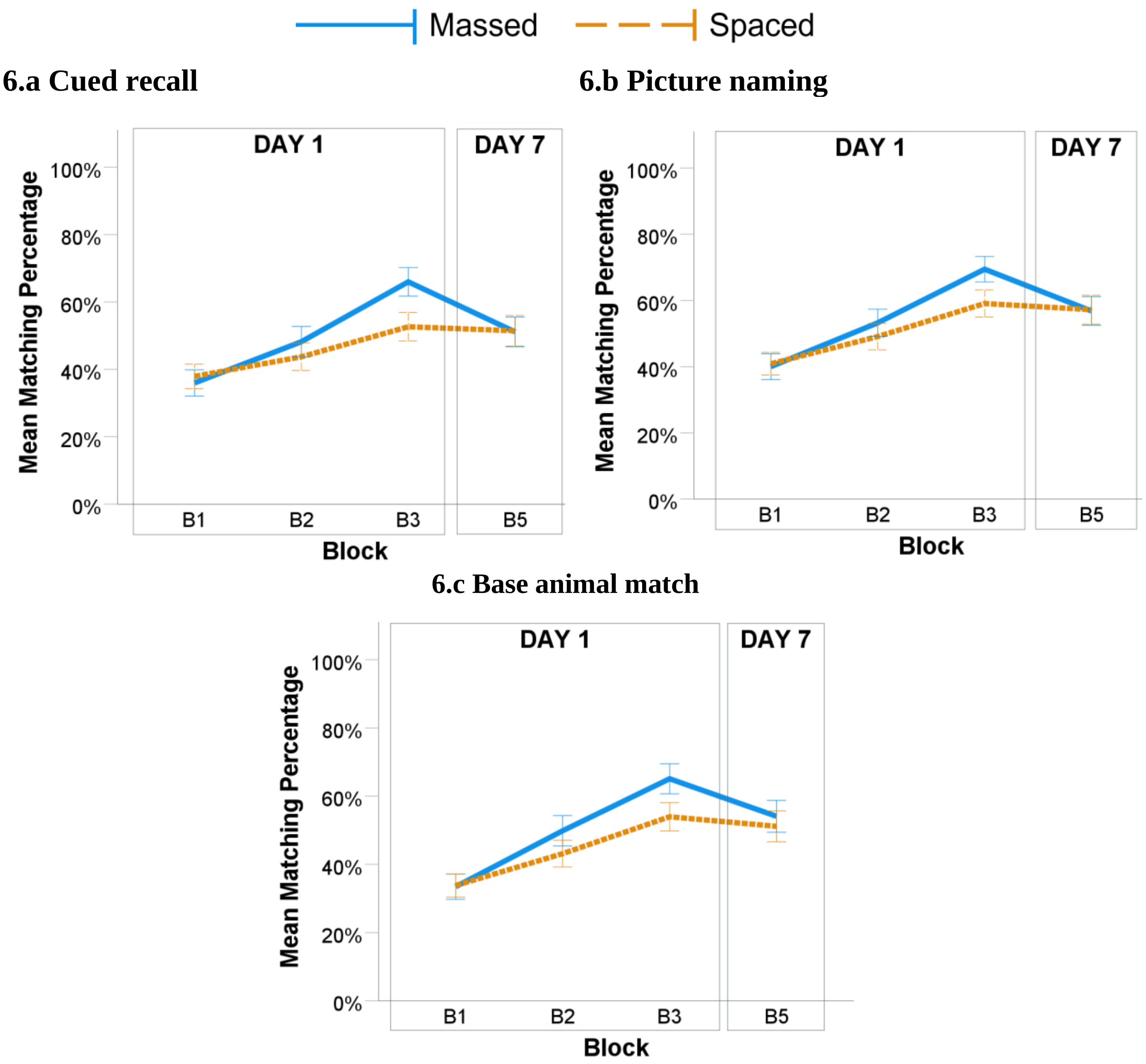
The matching percentage was calculated the same way as in Experiment 1. To test changes in performance across the three retrieval blocks on Day 1, we ran a 3(Block: B1, B2, B3) x 2(Schedule: Massed, Spaced) Bayesian repeated-measures ANOVA with each participant's average matching percentage as the dependent variable. To measure long-term changes in performance from Day 1 to Day 7, we conducted a 2(Block: B3, B5) x 2(Schedule: Massed, Spaced) Bayesian repeated-measures ANOVA on the three tasks.

### 2.5.2 Experiment 3 Results

One participant in the Massed schedule was identified as an outlier in the cued recall task, and their data was removed from relevant analyses.

#### *2.5.2.1 The effect of within-day spaced (or massed) retrieval practice with feedback before sleep (B1-3)*

The first analysis aimed to determine if overall performance improved from B1 to 3 and if there was an initial massed benefit (replicating the Day 1 results in Experiment 2). A 3(Block 1-3)x2(Schedule) Bayesian repeated-measures ANOVA was conducted with the average matching percentage as the dependent variable. There was overwhelming support for changes across Blocks for all three tests (cued recall $BF_{10}$=2.8e+19, error%=0.63; picture naming $BF_{10}$=1.9e+16, error%=1.65; base animal match $BF_{10}$=9.3e+19, error%=0.67). Figure 6 illustrates that all participants improved their performance across the three blocks on Day 1. There was weak to no evidence for the null hypothesis in the main effect of Schedule (cued recall $BF_{10}$=0.49, error%=0.55; picture naming $BF_{10}$=0.42, error%=1.03; base animal match $BF_{10}$=0.50, error%=1.69). However, there was moderate to overwhelming support for Block*Schedule for cued recall and base animal matching (cued recall $BF_{incl}$=533.3; base animal match $BF_{incl}$=3.18), but not for picture naming ($BF_{incl}$=1.82). Figure 6 shows steeper improvements in the Massed schedule than the Spaced schedule, replicating Experiment 2 for two of the three tests.



6.a Cued recall

6.b Picture naming

6.c Base animal match

**Figure 6. Experiment 3 results.** Graphs showing the mean matching percentage for participants in the two schedules across blocks on Day 1 and Day 7 in the cued recall (**6.a**), picture naming (**6.b**) and base animal match (**6.c**) tasks. Error bars represent ± 1SE.

#### *2.5.2.2 The effect of within-day spaced retrieval practice with feedback on retention of novel animal names one week after initial exposure (B3-5)*

The second set of analyses examined if within-day spaced retrieval practice resulted in a longer-term benefit in the memory performance. Again, the average matching percentage was used as the dependent variable in a 2(Block3, 5)x2(Schedule) Bayesian repeated-measures ANOVA. The analysis revealed strong and very strong support for a main effect of Block (cued recall $BF_{10}$=30.98, error%=2.21; picture naming $BF_{10}$=388.9, error%=4.91; base animal match $BF_{10}$=328.28, error%=2.50). In contrast to Experiment 2, Figure 6 shows overall forgetting from B3 to B5. There was no support for a main effect of Schedule (cued recall $BF_{10}$=0.60, error%=0.71; picture naming $BF_{10}$=0.57, error%=0.96; base animal match $BF_{10}$=0.68, error%=1.05). Nevertheless, there was strong support for a Block*Schedule interaction for both cued recall ($BF_{incl}$=18.78) and picture naming ($BF_{incl}$=33.93) and moderate support for base animal matching ($BF_{incl}$=3.94). Overall, as seen in Figure 6, participants undergoing massed retrieval practice showed forgetting from Day 1 to Day 7, whereas participants practising spaced retrieval maintained their performance from Day 1 (B3) to Day 7 (B5). Despite the greater level of forgetting in Massed, both schedules performed similarly on Day 7; therefore, there was no spacing benefit in long-term measures.

#### *2.5.2.3 SART and KSS*

The average reaction time of accurate hits and number of false hits (i.e. response when no response was required, when presented with "*X*") in the SART task was calculated for each participant to indicate levels of sustained attention at the final block of Days 1 and 7.

Comparing the average reaction time between the retrieval schedules in a 2(Block) x 2(Schedule) Bayesian repeated-measures ANOVA revealed no support for the Schedule variable ($BF_{10}$=0.52, error%=1.41) but support for the null hypothesis in Block*Schedule ($BF_{incl}$=0.39). The same pattern was found for the number of false hits, with no support for Schedule (BF10=0.34, error%=0.76) but weak support for the null hypothesis for Block*Schedule ($BF_{incl}$=0.31). These results suggest that levels of sustained attention did not differ between participants depending on their initial retrieval schedule, so the performance differences cannot be attributed to this.

For the KSS data, there was strong support for a null effect of Block ($BF_{10}$=0.03, error%=2.94) and weak support in the Schedule variable ($BF_{10}$=0.39, error%=0.54). There was weak support for a Block*Schedule interaction ($BF_{incl}$=3.08), with the Massed participants reporting higher sleepiness than the Spaced participants (by 0.6 points). However, the mean difference was very small (i.e., <1 point on the scale) and thus unlikely to fully account for the performance differences in the memory tests.

#### *2.5.2.4 Exploratory cross-experiment comparisons: Experiments 2 and 3*

To directly examine whether changes from Day 1 to Day 7 differ depending on if a post-sleep retrieval opportunity was given (i.e. including or excluding B4), we compared Experiment 2 (including B4) and Experiment 3 (excluding B4). Specifically, we conducted a 2(Experiment: Experiment 2, Experiment 3) x 2(Block: Penultimate block, Final block) x 2(Schedule: Massed, Spaced) Bayesian repeated-measures ANOVA, using the average matching percentage as the dependent variable in all tasks. To clarify, we compared performance from the penultimate block (i.e. Block 4 in Experiment 2 and Block 3 in Experiment 3) with a performance at the final retrieval block on Day 7 (i.e. B5 in both Experiments). In addition, we collapsed the data from participants in the Massed schedules in Experiment 2 to form a single Massed condition to be compared with the Spaced condition. This resulted in uneven sample sizes: Massed n

Experiment 2 = 60, Massed n Experiment 3 = 29, Spaced n Experiment 2 = 30, Spaced n Experiment 3 = 30. However, Bayesian analysis provides reasonable control for uneven sample sizes and significantly reduces the probability of Type I errors (Kelter, 2020).

Across all tasks, the analysis revealed support for an effect of Experiment (cued recall $BF_{10}$=356.2, error%=1.36; picture naming $BF_{10}$=1629.0, error%=2.10; base animal match $BF_{10}$=20277.9, error%=2.01), Block (cued recall $BF_{10}$=8.12, error%=2.57; picture naming $BF_{10}$=16577.4, error%=1.15; base animal match $BF_{10}$=712.1, error%=0.98), and some support for an effect of Schedule (cued recall $BF_{10}$=9.62, error%=4.83; picture naming $BF_{10}$=3.68, error%=3.95; base animal match $BF_{10}$=4.74, error%=4.14). Figure 7 shows that performance was higher for Experiment 2 than in Experiment 3, which was not surprising given the additional retrieval practice opportunity on Day 2 in the former. Overall performance decreased from the penultimate retrieval practice block to the final block, and there was also some indication that the Massed schedule performed higher than the Spaced overall. Importantly, there was strong support for an Experiment*Block*Schedule interaction for all tests (cued recall $BF_{incl}$=13.85; picture naming $BF_{incl}$=209.9; base animal match $BF_{incl}$=72.8). As seen in Figure 7, if no post-sleep retrieval opportunity was given on Day 2, participants in the Massed retrieval schedule showed forgetting. However, if the initial retrieval was spaced or an additional retrieval opportunity was given on Day 2, there were no forgetting and performance levels were maintained. This suggests that spaced retrieval practice opportunities (whether they are on the first day of learning or after a sleep period) safeguard against long-term forgetting. That is, the spaced group showed similar forgetting rates regardless of whether a post-sleep retrieval block was included, suggesting that the memory traces formed during within-day spaced retrieval opportunities were sufficiently protected from forgetting before sleep.

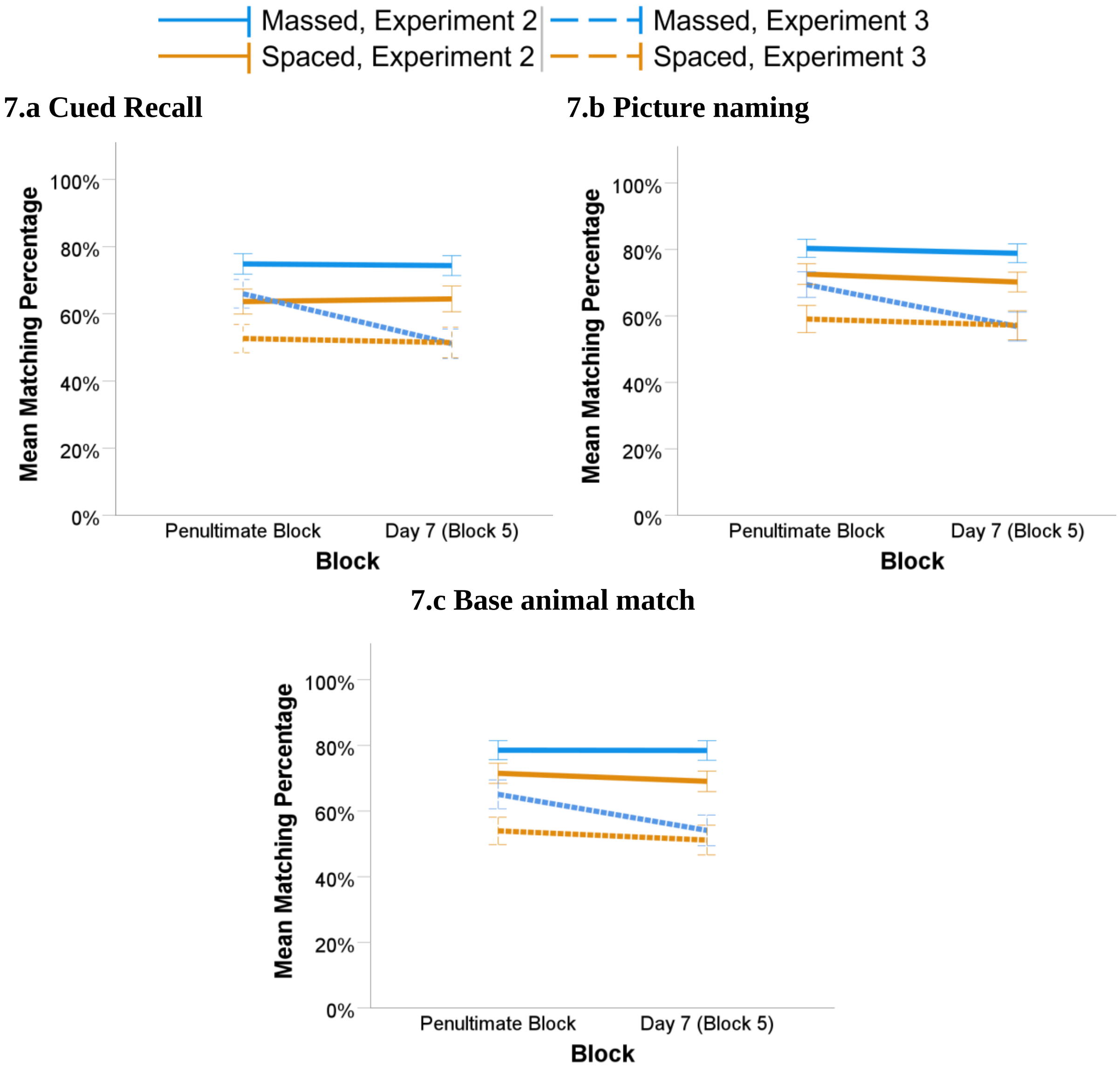


**Figure 7. Results of Experiment 2 and 3 comparison analysis.** Data shows changes in mean matching percentage performance in cued recall (**7.a**), picture naming (**7.b**), and base animal match (**7.c**) from the penultimate retrieval block (i.e. B4 in Experiment 2 and B3 in Experiment 3) to the final retrieval block (B5) on Day 7. Solid lines represent Experiment 2; dashed lines represent Experiment 3. Blue lines represent Massed Schedules; orange lines represent Spaced Schedules. Error bars represent ± 1SE.

### 2.5.3 Experiment 3 Discussion

Experiment 3 retained the use of feedback and examined whether within-day spaced retrieval practice would benefit the recall of newly learned words one week after learning compared to massed retrieval practice. Crucially, we omitted the day 2 (post-sleep) retrieval opportunity for all schedule conditions.

We confirmed the hypothesis regarding changes on Day 1: All participants improved their retrieval performance over tests, with those in the massed schedule again improving at a steeper rate than those in the spaced schedule in the cued recall and base animal match tests (as in Experiment 2). However, we still did not observe a spacing benefit on Day 7: participants in the Spaced schedule showed less forgetting than participants in the Massed schedule from Day 1 to Day 7, but they did not outperform participants in the Massed schedule on Day 7. Therefore, while initial improvements in recall levels were steeper in the Massed schedule than in Spaced, the lack of additional post-sleep retrieval attempts (arguably needed to trigger reconsolidation; Smith & Scarf, 2017) resulted in forgetting, lowering final performance one week later. This was further supported by the cross-experiment comparison, where it was clear that the reconsolidation opportunity on Day 2 protected the memories from Day 1 in the Massed schedule.

These results also suggested that initially practising retrieval of new material spaced over a day may exert protection from long-term forgetting. Thus, it follows that a spacing benefit may have emerged over a more extended period (e.g. one month) if the steeper rates of forgetting continued under conditions of initial massed retrieval practice. However, this is speculative, and our findings yet clearly contrast previous findings of a spacing benefit after shorter periods (e.g. two days Karpicke & Roediger, 2007; four days Bloom & Shuell, 1981).

## 2.6 General discussion

Three experiments investigated whether spacing retrieval practice over a day (relative to massing retrieval practice in a single session) benefits adults' new word learning before and after overnight sleep and one week later. In Experiment 1, no feedback was given following each retrieval practice, allowing us to determine whether the act of retrieval practice alone (as opposed to further exposure) influences memory performance. Under these conditions, there were only minor improvements across blocks, and the spacing of retrieval practice blocks across Day 1 (as opposed to a single massed session) did not lead to benefits in memory performance at any point (i.e., on Day 1, 2 or 7). Thus, in the absence of feedback, within-day spaced and massed retrieval practice leads to minor improvements in memory, but spacing did help or hinder memory performance compared to massed retrieval practice. Importantly, this severely questions the theoretical explanations of the benefits of spaced retrieval practice focused on the mechanisms of retrieval practice itself. Experiment 2 included feedback in the form of the correct answer in the retrieval practice blocks on Day 1. This resulted in greater gains over the Day 1 retrieval blocks and an initial benefit for massed (relative to spaced) retrieval practice. However, whilst participants who received spaced retrieval practice showed larger gains in performance from Day 1 to Day 2 (compared to the massed schedule), reducing their Day 1 deficit, this only worked to bring their performance in line with the massed schedule on Day 7. Finally, Experiment 3 was identical to Experiment 2 with the critical exception that we excluded the Day 2 tests to examine whether this provided a spaced retrieval opportunity for both schedules, thereby masking any delayed benefits of spaced retrieval at the one-week test in Experiment 2. Here, we replicated the initial massed benefit, but the two schedules again remained equivalent one week later, owing to greater forgetting for massed retrieval practice than for spaced (with maintained performance between Day 1 and Day 7 for the latter).

Across all experiments, at no point did the spaced schedule outperform the massed schedule on Day 7. These data fall in stark contrast to theories of spaced retrieval practice (e.g.

Maddox, 2016; Pyc & Rawson, 2009), with the results of Experiment 1 (without feedback to isolate retrieval practice from additional learning opportunities) lending particularly strong counter-evidence. It should be noted that a spacing benefit has been observed previously without feedback in different circumstances (e.g. when items are spaced between interleaved items within a single session; Karpicke & Roediger, 2007; Cull, 2000) or with/without feedback when initial learning levels are closer to ceiling (e.g. Bell et al., 2014). However, spaced retrieval practice alone (or in combination with feedback) was not enough to elicit a benefit in word learning here.

Instead, there was only nuanced evidence of a spaced retrieval effect, and only in the presence of feedback. Specifically, we observed (i) increased overnight improvements following spaced retrieval practice in Experiment 2 and (ii) reduced forgetting without additional reconsolidation opportunities in Experiment 3. Instead of providing clear evidence for a spaced retrieval practice benefit in word learning, the findings point to two factors that increase the effectiveness of retrieval practice in word learning: feedback and consolidation/reconsolidation opportunities.

### 2.6.1 Theoretical implications

#### *2.6.1.1 The role of feedback in spaced versus massed retrieval practice*

A key take-home message from the current study is that feedback influences the effects of retrieval practice. Whilst there were some modest improvements across blocks when feedback was excluded (i.e., in Experiment 1), these improvements were far greater when feedback was provided (i.e., in Experiments 2 and 3). This suggests that feedback was crucial for retrieval practice to aid learning processes, at least under the present conditions (for example, when initial exposure levels are low).

As mentioned in Experiment 1 Discussion (Section 2.3.3), the amount of initial exposure could explain why we did not observe similar results as other studies. There have been several suggestions that the optimal schedule of retrieval practice depends on the strength of memory at initial encoding (e.g. Latimier, Peyre & Ramus, 2021; Pavlik & Anderson, 2008; Raaijmakers, 2003). In particular, if initial encoding strength is high, retrieval attempts can be delayed by a temporal space, but a short delay is preferable if the initial encoding is limited (Latimier, Peyre & Ramus, 2021). This was reflected in the results of Experiments 1 and 2, as initial encoding strength was limited and performance on Day 1 benefited from massed retrieval practice. This could be because at a spaced retrieval attempt, weakly encoded memory risk falling below the threshold for successful retrieval, resulting in incomplete or incorrect retrieval. It has been found that retrieval practice can strengthen incorrect retrieval attempts, resulting in false memories (Zhuang et al., 2021; McDermott, 2006; Roediger, Jacoby & McDermott, 1996). By providing feedback, especially in *spaced* retrieval practice, we can combine the beneficial effects of spaced learning and spaced retrieval practice, reducing the risk of creating incorrect/false memories, even when initial encoding strength is low.

Further, there was evidence here that feedback affects the effectiveness of retrieval practice *schedules* on the retention of newly learned words. Support for this comes from the strong null effects of Schedule in Experiment 1 (without feedback), which contrasts with the strong Schedule effects in Experiments 2 and 3 (with feedback). Without feedback, spaced and massed retrieval practice performed similarly and showed only minor improvements over blocks. With feedback, the Massed schedules consistently performed better than Spaced before sleep, but with some evidence of spaced retrieval practice leading to greater “catch-up” overnight and less forgetting. A potential explanation for why the massed schedules improved at a steeper rate than Spaced on Day 1 when feedback was included could be linked to less time passing between retrieval attempts. Following contextual variability theories, the feedback could act as an overlapping cue, aiding retrieval in the massed blocks. In addition, retrieval would be easier as the newly formed memory traces could be reactivated through the feedback, so subsequent retrieval could rely on both short- and long-term memory stores. These accounts could explain why we observed steeper improvement rates for the massed schedules compared to spaced on Day 1. However, this would lead us to expect more rapid forgetting from massed retrieval practice after Day 1, resulting in a spacing benefit one week after initial exposure. This was not the observed pattern in Experiment 2, and while we saw an indication of more rapid massed forgetting in Experiment 3, there was still no evidence of a spacing benefit. Thus, these theories cannot explain why we did not see a spacing benefit.

In summary, our findings suggest that feedback is an integral part of ensuring retrieval practice is a successful learning tool and important for eliciting early benefits of massed (relative to spaced) scheduling. However, there is no clear evidence here that feedback is integral to eliciting a *spacing benefit*.

#### *2.6.1.2 The role of consolidation and reconsolidation in spaced retrieval practice*

The absence of an overall advantage in memory 24 hours and one week after new word learning when retrieval practice was spaced relative to massed suggests that a spacing benefit is not guaranteed when no intervening sleep occurs between the spaced retrieval practice opportunities. According to Smith and Scarf (2017), reconsolidation during spaced retrieval practice can aid later retrieval by slowing forgetting, consistent with many spaced retrieval practice studies that span at least one sleep period (e.g. Bell et al., 2014).

The steeper overnight improvement for the spaced schedule makes it tempting to conclude that overnight consolidation and reconsolidation processes are more effective following spaced than massed retrieval practice. However, potentially a more plausible explanation could be linked to the performance levels before sleep (i.e., block 3 by the end of Day 1). Overnight consolidation can benefit (i.e. aid performance after sleep) both strongly and weakly encoded memories, but overnight changes can be more easily detected for weak memories as there is more room for improvement (e.g. Petzka et al. 2020). Because participants in the spaced schedule performed lower than those in massed schedules by the end of day 1 (i.e., block 3), there was more room for potential improvement from consolidation and reconsolidation opportunities. To examine this suggestion, we ran additional correlational analyses between performance in the final session of Day 1 and the overnight change (overnight change = mean matching percentage at Block 4 - mean matching percentage at Block 3; see Appendix A2). These analyses revealed a negative correlation, suggesting that lower performance by the end of day 1 related to a bigger change in performance overnight. Thus, performance levels before sleep may affect benefits from overnight consolidation (as per Petzka et al., 2020) and the effectiveness of reconsolidation after a period of sleep.

Next, we will focus on performance changes from the end of Day 1 to one week later. In Experiments 1 and 2, we observed that all participants maintained their performance from Day 2 to Day 7, indicating that the reconsolidation opportunity on Day 2 allowed participants to maintain their memory representations of the animal names (supporting Smith & Scarf's argument, 2017). In Experiment 3, we observed more forgetting from Day 1 to Day 7 if the initial retrieval practice was massed rather than spaced. While this is somewhat in line with extant literature often reporting a spacing benefit after more extensive forgetting from massed retrieval practice, we cannot claim that our findings indicate a spacing benefit. Participants in the spaced schedule did maintain their performance better than those in the massed schedule, but they never outperformed the massed schedules. Hence, we found evidence that within-day spaced retrieval practice protected memories from forgetting, but it did not result in a spacing *benefit*.

A possibility for the reduced level of forgetting from within-day spaced retrieval practice could be that the retrieval practice opportunities encouraged online consolidation of the animal names (Antony et al., 2017). Online consolidation is consolidation that occurs outside of a sleep period. For example, through concurrent activation of hippocampal and neocortical regions during wake, allowing communication between the areas, strengthening the memory traces. Antony and colleagues argued that retrieval practice is an effective learning tool because it can encourage this type of online consolidation, resulting in more effective embedding of the novel memories in neocortical regions. Although this argument on its own is not sufficient to explain why we saw performance differences between schedules in Experiment 3 (as both spaced and massed schedules practiced retrieval), it could be extended to include a certain threshold of difficulty for effective consolidation through retrieval practice. For example, since less time passed between massed retrieval attempts, retrieval would be easier and could, speculatively, rely more on local short-term memory processes rather than coordinated communication between cortical regions. Thus, this might not have been sufficient to encourage the consolidation processes as suggested by Antony et al. (2017). Instead, retrieval attempts would be harder in the spaced schedule, potentially encouraging online consolidation, resulting in better established memory traces, explaining the maintained performance in the Day 7 measures. This explanation fits with the findings of Experiment 3 and could extend Antony et al’s suggestion to include the caveat that retrieval needs to be more challenging for effective consolidation to occur without sleep or reconsolidation processes aiding it.

In short, while Experiment 1 and 2 provides evidence of the importance of feedback and consolidation/reconsolidation processes in spaced retrieval practice (i.e. maintained performance after retrieval blocks spanning two days, regardless of Day 1 schedule), Experiment 3 suggests that within-day spaced retrieval practice protects memories from forgetting more than massed retrieval practice. Thus, the lower levels of forgetting in the spaced condition observed in Experiment 3, and the steeper overnight improvements in Experiment 2, indicate that both within-day spaced retrieval practice with feedback and reconsolidation play a role in eliciting the effects of spaced retrieval practice and should be included in theoretical accounts moving forward.

### 2.6.2 Educational implications

Spaced retrieval practice is often highlighted as a successful example of neuroscience applied in real-world educational settings. However, there is substantial doubt about when a spacing effect occurs and when it does not. As shown clearly in the current experiments, spaced retrieval practice does not always result in a spacing *benefit* in long-term measures. In fact, the findings of Experiments 2 and 3 suggest that Massed retrieval practice can be more beneficial if novel words only need to be remembered for a relatively short time (e.g. 24 hours, up to one week). However, it is unclear whether this remains in longer-term measures (i.e. longer than one week). The effects of spaced retrieval practice appear to be nuanced and multifaceted and we still do not clearly understand the boundary conditions of spaced retrieval practice (i.e. when it is effective and when it is not). Our study suggests that feedback and consolidation/reconsolidation are two crucial topics for further exploration. However, until we have a clearer picture of what aspects are needed for a spacing benefit to emerge (e.g., feedback, spacing, consolidation/reconsolidation), spaced retrieval practice should only be cautiously advised in educational settings.

Further, as shown in the current experiments in combination with previous studies, sleep may play a role in (1) bolstering the effects of initial spaced retrieval practice and (2) protecting from forgetting following massed retrieval practice; as such, spacing out retrieval practice over a sleep period may be particularly advantageous. However, the current literature does not have enough evidence to build this argument further or whether this can be applied to all populations, regardless of sleep architecture. For example, it could be speculated that populations with differences in sleep architecture could benefit from spaced retrieval practice differently. Some studies have looked at this in young and old adults (e.g. Coane, 2013; Logan & Balota, 2008), but surprisingly, no studies to our knowledge have directly compared adults and children. Children have shown greater benefits than adults from overnight consolidation when learning novel words (e.g. James et al., 2019; Wilhelm, 2014), which have been partially attributed to increased activation supporting sleep-based consolidation processes (Ohayon et al., 2004). Thus, if consolidation and reconsolidation are important in eliciting an effect from spaced retrieval practice, we must close the empirical gap of studies directly comparing spaced retrieval practice in adults and children.

## 2.7 Conclusion

Overall, within-day spaced retrieval practice does not solely seem to emerge from harder retrieval or varied context. Instead, our findings support the argument that feedback is crucial for the effectiveness of spaced retrieval practice, and consolidation and reconsolidation opportunities are part of the explanation. We did not observe an overall longer-term advantage in memory performance one week after exposure for spaced retrieval practice relative to massed cramming. Whilst we did observe tentative evidence that within-day spacing may protect against forgetting, it remains to be seen whether a spacing advantage emerges over a longer period. Considering this, the current theories of spaced retrieval practice appear to be insufficient, and the use of spaced retrieval practice in educational settings should be revisited, not least to ascertain whether our conclusions can be applied at different points of development and different levels of learning.

# Chapter 3:

## Does within-day spaced retrieval practice result in a spacing benefit in school-aged children? A classroom-based study

The stimuli, data and analysis outputs for Experiments 4-5 are available at the Open Science Framework: https://osf.io/tvcm9/

The pre-registration for Experiment 5 is available at the Open Science Framework: https://osf.io/s4azr

## 3.1 Abstract

Spaced retrieval practice has been promoted as an effective learning strategy to be used in schools to promote long-term retention of memories. Children have been shown to benefit from and rely on sleep-based consolidation to a great extent in word learning, potentially affecting the outcome of spaced retrieval practice that spans across sleep. However, few studies have examined the effects of spaced retrieval practice without the interference of intervening sleep between retrieval occasions, especially in classroom settings. We conducted two experiments where children learned and practised retrieval of novel animal names in two sessions spaced 3 hours apart or in a single massed session. In Experiment 4, children completed a follow-up session 24 hours later but not in Experiment 5. One week later, all children completed a final long-term retention test. We found evidence that children could learn more animal names if practising retrieval in a single massed session than in two spaced sessions, contradicting previous findings of an early emerging spacing benefit. Interestingly, we also did not observe a spacing benefit one week later. Instead, if children were given a retrieval opportunity 24h after initial exposure (i.e., after one night's sleep), performance continued to improve, resulting in the highest retention one week later, regardless of the initial retrieval schedule. However, when no retrieval opportunity was given the day after initial training, only children in the spaced schedule continued to improve, while those in the massed schedule maintained their performance. Tentatively, these findings lend nuanced evidence for a spacing benefit. However, current theories cannot explain our findings, highlighting the need for continued research of spaced retrieval practice, especially in schools, before promoting it as a viable method in educational settings.

## 3.2 Introduction

Spaced retrieval practice has been highlighted as one of the most promising areas of research in promoting long-term retention of novel memories, including the retention of new vocabulary (Kornmeier, Sosic-Vasic & Joos, 2022; Dunlosky et al., 2013). Most research in spaced retrieval practice draws on adult populations, but the recommendation for its use has been extended to various populations and materials. For example, spaced retrieval practice has been encouraged as an optimal learning method in school settings, both in general and in the focus of word learning (Surma, Vanhoyweghen, Camp & Kirschner, 2018). However, relatively few studies have examined whether spaced retrieval practice is beneficial for long-term retention in child populations, especially in classroom settings (in contrast to more traditional laboratory settings). This study comprises two experiments examining the frequently cited claim that spaced retrieval practice can lead to long-term memory benefits in the context of word learning. Crucially, the experiments were conducted with school-aged children, allowing us to expand theoretical knowledge and examine practical implications from a developmental perspective. The focus here is on word learning, a critical process supported by our memory systems (Davis & Gaskell, 2009), culminating in our vocabulary knowledge as understanding the mechanisms of word learning and maximising long-term retention of new words during childhood is key to informing effective educational practice.

### 3.2.1 Word learning in development

As outlined in Chapter 1, adults' novel word learning undergoes a consolidation process, incorporating novel words into long-term lexical networks and facilitating effective use in everyday language. The complementary learning systems framework provides one perspective on how this process may unfold (McClelland et al., 1995; O'Reilly & Rudy, 2000; O'Reilly et al., 2014). The CLS framework argues that memory relies on two learning and memory systems that complement each other, i.e. hippocampal regions support rapid learning of sparse memory traces while neocortical regions support slower learning, providing rich and long-lasting memory traces. Consolidation occurs through coordinated communication between the cortical regions, often occurring during sleep periods (i.e., sleep-based consolidation), reducing dependency on hippocampal regions and enhancing activation in neocortical regions in later retrieval of memories (O'Reilly, Bhattacharyya, Howard & Ketz, 2014; Tamminen et al., 2010; Davis & Gaskell, 2009). When applied to word learning (Dumay & Gaskell, 2007), this framework has received support from findings in the adult literature (e.g., Dumay & Gaskell, 2012; Tamminen et al., 2010; Davis & Gaskell, 2009).

Crucially, similar findings have been demonstrated in children (e.g., Henderson, Devine, Weighall & Gaskell, 2015; Henderson, Weighall, Brown & Gaskell, 2012; Brown, Weighall, Henderson & Gaskell, 2012; Gais, Lucas & Born, 2006). For example, Henderson et al. (2012) taught children aged 7-12 years novel words either in the AM or the PM. At a test administered 12 hours later (i.e., after a night of sleep for the PM group but a day of wakefulness for the AM group), the PM group could recall significantly more words than those in AM group and showed signs of lexical integration (i.e., demonstrating lexical competition effects between existing words, e.g., "*dolphin*", and new competitor words, e.g., "*dolpheg*"). In contrast, the AM condition did not show improvements in recall or signs of lexical integration until after a sleep period (i.e., at a 24-hour test). Additionally, while a full night’s sleep can facilitate word learning in children, a shorter sleep period, such as a daytime nap, has also been found to enhance word learning in children as young as three years (Williams & Horst, 2014). These findings suggest that children, like adults, can utilise sleep to consolidate novel words, supporting the importance of consolidation and the CLS model in word learning in children.

Interestingly, however, there is accumulating evidence of subtle developmental differences in the magnitude of overnight consolidation effects when direct comparisons are made between adults and children. In general, the child literature reports a pattern of *improvements* in retention levels after sleep, whilst adults tend to show reduced forgetting and/or smaller post-sleep improvements. For example, Brown et al. (2012) taught children (aged 6-8) novel non-words (e.g. *biscal*) that had a corresponding neighbour word already in the mental lexicon (e.g. *biscuit*). The children learned and were tested on (using cued recall) the novel words either in the morning and afternoon (e.g. 9 am and 1 pm; *no sleep group*) or in the morning of two consecutive days (e.g. 11 am on day 1 and 11 am on day 2; *sleep group*). When tested on recall of the novel words, the sleep group showed better performance than the wake group, demonstrating how sleep can improve children's performance even without further exposure to the novel words. Similar findings have been reported in older children aged 7-12 (Henderson et al., 2012), where both recognition and recall performance of non-words improved after a period of sleep, not after wake. Studies that have directly compared children and adults also report larger overnight improvements in recall of novel words in children (e.g., James, Gaskell & Henderson, 2018; Weighall et al., 2017) and precisely how sleep-based benefits may differ between children and adults will be further considered in Chapter 4. Furthermore, it has also been found that the size of the overnight change in word recall is positively associated with levels of prior knowledge in children (e.g., James et al., 2017). Specifically, in word-learning tasks, higher levels of prior vocabulary knowledge are correlated with greater overnight gains in recall (e.g. James et al., 2017; Henderson et al., 2015), particularly when children are learning new words more implicitly across multiple story contexts (Henderson & James, 2018). Collectively, these findings emphasize the role of sleep in the long-term retention of new words in development and that variables intrinsic to the child (e.g., vocabulary knowledge) can influence this.

### 3.2.2 Spaced retrieval practice in development

In addition to processes of sleep-based consolidation and the intrinsic vocabulary knowledge that a child brings to the task, the extrinsic conditions of initial encoding have also been found to influence the long-term recall of novel words in children. Indeed, spaced retrieval practice is often recommended as an optimal learning condition to promote long-term retention of new information in adults and children. A frequently cited study by Bloom and Shuell (1981) taught French words to high-school children who practised retrieving these in one 30-minute session on one day (the *massed* condition) or three 10-minute sessions on three consecutive days (the *spaced* condition). When tested on recall four days later, the spaced group could recall significantly more words (35% more words) than the massed group.

Since this seminal study, the developmental literature on spaced retrieval practice has somewhat expanded, but the empirical gap in studies with children remains. An example of a study exploring spaced retrieval practice in children is by Petersen-Brown and colleagues (2019), who taught children in grade 3 and 4 maths vocabulary on Day 1 and used flashcard type retrieval practice with corrective feedback in a *fixed spaced* schedule (Day 7, 14, and 21), an *expanding spaced* schedule (Day 2, 9, and 21) or a *massed* schedule (all on Day 1, in the same session). A long-term retention test using the same flashcard task was conducted seven days after the final retrieval practice session. While the massed group performed better than the two spaced groups throughout the three retrieval practice sessions, the spaced group outperformed massed one week later. This was because the massed group showed more forgetting than the two spaced groups, who instead improved their performance. Overall, the consensus is that spaced retrieval practice results in better long-term recall performance in children, similarly to adults.

#### *3.2.2.1 Applying theories of spaced retrieval practice to developmental data*

There are several theories of why spaced retrieval practice enhances long-term memory, all of which have been developed from adult data (as outlined in more detail in Chapter 1). In short, the c*ontextual variability theories* suggest that spaced retrieval practice allows for a greater variety of contextual cues to be integrated with novel memory traces, creating more cues to aid long-term retrieval and leading to better longer-term memory performance (Estes, 1955; Pashler et al., 2009; Maddox, 2016). Alternatively, Pyc and Rawson (2009) proposed the *retrieval effort hypothesis*, arguing that complex retrieval leads to better long-term retention than easier retrieval because more challenging retrieval activates memory traces that decay slower (Pavlik & Anderson, 2008). Finally, considering the importance of sleep in word learning, Smith and Scarf (2017) proposed a *reconsolidation account*. They suggest that partially consolidated memories undergo a reconsolidation process when retrieval practice is spaced over a sleep period. That is, a partially consolidated memory trace is reactivated (e.g. through retrieval), placing it in a fragile state that elicits further consolidation during subsequent sleep. Thus, learning novel words through retrieval practice events that span periods of sleep may benefit from both consolidation and reconsolidation, resulting in more established memory traces and better long-term retention.

Crucially, however, these theories cannot explain the lack of spacing benefits in adults after within-day spaced retrieval practice, as demonstrated in Chapter 2. Indeed, Experiments

1-3 showed that adults who learned novel animal names in repeated retrieval practice sessions did not benefit from sessions being spaced 2 hours apart when tested on memory recall one week later. This was despite performance initially adhering to the predicted pattern of contextual variability and retrieval difficulty theories (i.e., a benefit for recall of novel words following massed retrieval practice relative to spaced retrieval practice). Thus, these theories cannot explain why a spacing benefit only emerges in some circumstances.

A further and fundamental limitation of the spaced retrieval practice literature is that it is based mainly on adult data, placing constraints on the advance of theoretical accounts and practical application. For example, the contextual variability and retrieval difficulty theories struggle to explain why some studies have found an earlier emerging spacing benefit in children (i.e. a spacing benefit in the initial retrieval practice sessions, not just in long-term measures). For example, Ambridge et al. (2006) found that 4-year-old children learned and performed better when exposed to grammatical constructions in sessions spaced over five consecutive days than if all exposure were in a single, massed session. Similarly, Childers and Tomasello (2002) found that children could produce more novel words after spaced training than massed. The contextual variability and retrieval difficulty theories would expect an initial benefit from massed retrieval practice (due to more overlapping cues to aid initial retrieval and shorter time passing between retrieval attempts making it easier). Aligning with this, we found an initial benefit after massed retrieval practice before sleep in Experiments 2 and 3 in Chapter 2, a pattern often seen in the extant adult literature (e.g. Maddox, 2016; Delaney et al., 2010; Cepeda et al., 2006). This indicates that spaced retrieval practice is mainly beneficial for reducing forgetting in long-term measures, resulting in a benefit in later tests. Instead, child literature on spaced retrieval practice often reports both an initial and long-term benefit from spaced retrieval practice, suggesting that the spaces benefit initial learning and long-term retention (e.g. Ambridge et al., 2006; Childers & Tomasello, 2002). This pattern cannot be explained by retrieval effort or contextual variability theories, as these theories are based on adult findings and do not differentiate between developmental stages.

#### *3.2.2.2 Explaining spaced retrieval practice in children: Sleep and reconsolidation*

As suggested by Smith and Scarf (2017), considering the role of sleep and reconsolidation in spaced retrieval practice could provide a more inclusive theoretical account for why we can observe developmental differences in the spacing effect. Children can show sleep-based consolidation effects similarly to adults (some even argue bigger effects, e.g., James, Gaskell & Henderson, 2019; Wilhelm et al., 2014). According to their argument, overnight sleep could consolidate parts of words learned before sleep, allowing room for further learning in subsequent spaced retrieval practice sessions. In contrast, if retrieval practice is massed into a single session without intervening sleep, there might not be room for further improvements as no consolidation opportunities could help relieve the learning load. In addition to consolidation through sleep, while not discussed by Smith and Scarf, this argument would place great value on the inclusion of feedback in the retrieval practice sessions, as this would allow for further learning to take place in combination with the beneficial effects of retrieval practice. This is in line with our findings in adults, as discussed in Chapter 2, where excluding feedback in repeated retrieval practice did not improve performance, but including feedback did. As such, consolidation and reconsolidation could account for the spaced retrieval practice effects in children. The differences in sleep-based consolidation benefits could therefore explain why adults and children show benefits at different stages of the spacing process.

The above argument highlights the importance of differentiating between *sleep* and *no-sleep* spaced retrieval practice, especially when discussing the effect in children given that they can show enhanced benefits from sleep-based consolidation compared with adults (as mentioned briefly above and discussed in more detail in Chapter 1). Further, varying the number of intervening nights (i.e. sleep) between spaced sessions in word learning tasks could affect the later spacing benefits observed. For example, Schwartz and Terrell (1983) found that young children (0-3 years) who practised retrieving novel words over ten days could learn more than if the practice was spread out over five days. Subsequently, Childers and Tomasello (2002) reported that children could produce more words if they had practised retrieving the words in sessions spaced over four consecutive days (one short session/day) than over two consecutive days (one longer session/day). Thus, short sessions spaced over several days led to the most efficient learning in children. However, they also found that when four shorter practice sessions occurred, it did not matter if they were separated by 24 hours or three full days. This could suggest that the total time between sessions does not matter as much as having at least one intervening sleep period in children.

When examining the literature, it becomes evident that most studies reporting a spacing benefit in retrieval performance in word learning tasks have at least one sleep period between retrieval occasions (e.g., Petersen-Brown et al., 2019; Goossens et al., 2012; Sobel et al., 2011). Thus, the benefits from spaced retrieval practice could be attributed to sleep rather than the temporal space. For example, Sobel et al. (2011) taught children (10 years old) novel words and their definitions in one longer session (massed) or in two shorter sessions taking place one week apart (spaced). At a final recall test five weeks after the final spaced or massed practice occasion, the spaced group could recall three times more words and definitions than the massed group. In the spaced group, children had several opportunities for sleep-based consolidation to strengthen and enhance memory traces of the novel words in neocortical regions between practice sessions.

Thus, it is possible that in the second spaced session one week later, the already partially consolidated memories could have been reactivated and placed in a fragile state (i.e. *reconsolidation*; Smith & Scarf, 2017), potentially allowing further sleep-based consolidation in the following nights. However, the massed group might not have benefited from reconsolidation in the same way as there was no chance for the novel memory traces of the words to become consolidated between practice opportunities. That is not to say that the words learned in the massed manner did not benefit from sleep-based consolidation; there were several opportunities for sleep between the initial learning session and the final retention measure. However, the lack of *reconsolidation* opportunities could have resulted in more negligible improvements in memory traces and more forgetting in long-term measures. Thus, sleep-based consolidation and reconsolidation processes could potentially explain many reported patterns in children's spacing literature on word learning. It should be noted that Childers and Tomasello (2002) reported that the production of novel words was equal regardless of whether the practice opportunities were spaced by 24 hours or three days, as long as the sessions were short. So while sleep and reconsolidation can explain many overall patterns from spaced retrieval practice, it is essential to explore spaced retrieval practice without the influence of sleep to determine whether other factors play a role. Unfortunately, few studies have explored within-day spaced retrieval practice in word learning in children.

One exception is Seabrooke et al. (2005), who taught children in grade 1 reading skills (e.g. grapheme-phoneme correspondence and word reading) every day for two weeks, in a single 6-minute session (*massed*) or three 2-minute sessions spread out over the school day (*spaced*). After two weeks of intervention, the children who learned reading skills in the spaced sessions had improved significantly more than the children in the massed sessions. While this study does not explicitly look at word learning, and sessions were spaced over both within-day time and across days, their findings indicate that spacing short learning sessions within a single day can be more effective than massed sessions in enhancing children's reading skills. This raises the question of whether within-day spaced retrieval practice effectively promotes long-term retention of novel words in children (in contrast to adults, i.e., see Chapter 2). Indeed, based on Seabrooke et al., it seems that spaced retrieval practice without intervening sleep can be more effective than massed retrieval practice for children. However, this may only be the case if sufficient time passes between spaces.

This raises another limitation of the current literature on spaced retrieval practice, with studies spacing tests over minutes (e.g., Wegener et al., 2022) or days (e.g., Petersen-Brown et al., 2019) both being considered as examining spaced retrieval practice, despite these two cases most likely tapping into different underlying processes (i.e., especially given only the latter includes interleaving sleep). Furthermore, comparing studies using a space of minutes and days reveals a different pattern of benefits from spaced retrieval practice. For example, Petersen-Brown et al. (2019) found that a space of 7 days between practice occasions resulted in an initial massed benefit (measured immediately after the final spaced or massed session) but a later spacing benefit in long-term recall measures (measured seven days after the initial learning session). In contrast, Wegener and colleagues taught children (7-9 years) novel words (embedded in sentences) in spaced (2.5 minutes apart) or massed (back-to-back) sessions. Seven minutes after the final spaced or massed session, the spaced group could *recognise* more novel words, but there was no difference in cued *recall* measures. If Wegener and colleagues included a long-term measure after a more extended period (e.g., two days later), one could speculate that a more robust spacing benefit could have emerged (i.e., also for recall), similar to the findings in Petersen-Brown, due to more forgetting in the massed group[3]. However, to determine whether spaced retrieval practice benefits emerge without intervening sleep in children, it will be important to examine this directly.

Another limitation of the current literature is the lack of classroom/group-based testing. While lab-based or individual testing studies can provide crucial theoretical information when exploring educationally relevant topics, such as spaced retrieval practice, they do not consider real-life, practical aspects. For example, more complex methodologies used in spaced retrieval practice research cannot be implemented practically in classroom studies. For example, Petersen-Brown and colleagues conducted their study as one-on-one testing (i.e., children were taken out of their usual classroom activity to participate). While this is a valid method if a student requires individual tutoring, the findings' educational implications are reduced as most schools primarily practise classroom-based learning. In addition, if a participant in Petersen-Brown's study did not recall one of the words correctly during an initial retrieval attempt, the instructor repeated the query until the participant could successfully recall it. However, this kind of personalised learning schedule would be challenging to implement consistently in classrooms. Another limitation would be scheduling the retrieval practice sessions over a

[3] The differences in performance could alternatively be linked to the different types of tasks used in the two studies. Petersen-Brown used a cued recall task where participants recalled the novel words and definitions in a flashcard task. Instead, Wegener et al. used a dictation task where the novel words were read out, and participants had to spell the words from memory. Thus, the two tasks indicate different memory processes (i.e. semantic recall and orthographic memory) but are still classified as looking at word learning.

timeframe outside the traditional school day. As this type of schedule is not relevant in typical educational settings, findings from studies adopting this design should be applied with caution. Thus, using viable methods and appropriate timeframes within a typical school day is vital when planning research on spaced retrieval practice. There has been an increase in classroom-based studies exploring word learning through spaced retrieval practice in the past decades (e.g., Goossens et al., 2016; Goossens et al., 2012; Sobel et al., 2011), but the empirical gap remains, potentially reducing the credibility of the recommendations of spaced retrieval practice use in classroom-based teaching.

### 3.2.3 The current experiments

We addressed the limitations outlined above by conducting two experiments to determine whether within-day spaced retrieval practice was more beneficial than massed retrieval practice when children learn novel words in a classroom setting. Both Experiments had a priori hypotheses, with the methods and analyses for Experiment 5 preregistered at https://osf.io/s4azr [4]. These experiments followed the same materials and procedure as Experiments 2 and 3 in Chapter 2, as this would allow for a direct comparison of the effects of spaced retrieval practice between adults and children in Chapter 4. However, to accommodate the full retrieval schedules within a single school day, we adapted the schedules of the current experiments to only include two spaced blocks on Day 1 (instead of 3 in the adult experiments). We also reduced the number of animal names to learn to 12 (from 20) to avoid potential

[4] Note that the preregistration includes comparisons of child and adult data. Due to the structure of this thesis, these comparison will be outlined and discussed in Chapter 4.

performance at floor level, as children can have a lower capacity to learn novel words (e.g., Childers & Tomasello, 2002; Schwarts & Terrell, 1983).

The two experiments took place in schools in the Yorkshire (UK) area, where school-aged children were taught names (and associated photographic images) of real but rare animals (based on Fletcher et al., 2020) in their classrooms. Natural material was selected to be more pedagogically meaningful and motivating for participants to learn than compared to novel nonsense words (as are often used in studies of word learning, e.g., Dumay & Gaskell, 2007; Henderson et al., 2012). All experiments began with an initial exposure session, which was intentionally brief and intended to lay the foundation to observe change (i.e., improvements, maintenance, forgetting) over subsequent retrieval attempts. Following this, two Blocks of retrieval practice were administered (B1 and 2). On Day 1, these were either *massed* (all taking place back-to-back in a single session) or *spaced* (by around 3 hours). This design allowed all Blocks to occur within the same school day, without intervening sleep potentially affecting immediate and later retention of the novel words. By separating the Blocks by 3 hours, theoretically, retrieval should be more challenging (due to more time passing, allowing forgetting to occur; e.g., Pyc & Rawson, 2008) and/or the temporal context should be more varied than if retrieval practice is massed (e.g. Karpicke, Lehman & Aue, 2014). Thus, if we observe a spacing benefit on Day 1, this cannot be attributed to sleep-associated consolidation or reconsolidation processes. In Experiment 4, we included two Massed schedules to counteract time-of-day effects (e.g., Henderson et al., 2014): Massed AM completed their exposure/testing in the morning and Massed PM in the afternoon. To determine whether the initial spaced or massed retrieval practice affected performance after one night's sleep (e.g., whether sleep worked to enhance or reduce any benefit of initial spaced retrieval practice), Experiment 4 included a follow-up Block on Day 2 (B3). Finally, to determine whether any benefit of spaced retrieval practice is maintained in the longer-term (importantly, if spaced retrieval practice is to be recommended as a tractable learning aid), both experiments included a final Block one week after Day 1 (B4).

As in Experiments 1-3 in Chapter 2, we deployed three retrieval practice tests designed to activate phonological, orthographic, and semantic aspects of word-form memory. We measured participants' memory through written responses in all tests, capturing subtle changes in how memory traces matched the exposed form. The tests comprised: (i) a cued recall test, where participants were presented with the first two letters of a word and asked to provide the complete word, and (ii) a picture-naming test, where participants were presented with a picture and asked to provide the word semantically associated with it, and (iii) a base animal match test in which participants were asked to retrieve a newly learned word that was associated with a picture of a familiar animal (e.g., a picture of a hedgehog should elicit the written response "TENREC" - a small hedgehog with distinctive yellow and black stripes). The latter test was incorporated to encourage retrieval of the novel words in parallel with the retrieval of already known words with shared semantic properties. In addition, we included corrective feedback (i.e., the correct answer) in the base animal match test as we found that it was a crucial part of allowing further learning to occur in retrieval practice and, in extension, *spaced* retrieval practice (see Chapter 2). It was essential to allow further learning in the retrieval attempts as the initial exposure was limited, so to avoid performance at floor levels, we wanted to encourage further learning in the retrieval occasions while maintaining the beneficial effects of retrieval practice. Utilising these methods, Experiment 4 aimed to determine whether within-day spaced retrieval practice resulted in a benefit in word recall before and/or after sleep and one week after initial exposure in children, and Experiment 5 examined whether a post-sleep retrieval opportunity affected this.

## 3.3 Experiment 4

Experiment 4 examined whether a within-day spaced retrieval practice schedule leads to memory benefits before or after sleep and in longer-term measures (one week later) in tests designed to target the memory of novel words. School-aged children in Years 4 and 5 (aged 9-10) participated in a classroom setting, allowing us to get closer to examining whether the methods (and results) could apply to classroom teaching.

If sleep is not needed for a spacing benefit to emerge, we hypothesised that participants in the Spaced schedule would learn more words than participants in Massed on Day 1 due to children benefiting from spaced practice sessions (e.g., Childers & Tomasello, 2002). This would be manifested as a main effect of Schedule and potentially by a Block*Schedule interaction in all the three retrieval tasks (i.e. cued recall, picture naming, base animal match). Furthermore, focusing on overnight changes in memory performance in the three retrieval tests, we hypothesised that all children would show improved performance from Day 1 to Day 2 (as a consequence of both sleep-based consolidation and repeat testing/practice effects), but with the Spaced schedule maintaining their potential benefit from Day 1. Finally, if within-day spaced retrieval practice efficiently promotes a long-term benefit without the potential interference from intervening sleep-based consolidation, we hypothesised that the spaced retrieval practice schedule would result in better performance (i.e. more words recalled correctly) one week later (i.e. Day 7). This would again manifest as the main effect of Schedule and a Block*Schedule interaction to indicate that the initial retrieval practice schedule affected the retention of novel words differently over time. Specifically, based on previous studies with child populations (e.g., Wegener et al., 2022; Petersen-Brown et el., 2019; Goossens et al., 2012, Sobel et al., 2011), we expected children in the Massed schedules to display more forgetting than those in Spaced, who would maintain or improve performance from Days 1 to 2 and 7 in the three retrieval tests.

### 3.3.1 Experiment 4 Methods

#### *3.3.1.1 Participants*

Three Year 5 classes were recruited from a primary school in Yorkshire, UK. There were a total of 74 children (age range 9-10 years; mean age = 10 years and 1 month), and each class was assigned to a different retrieval practice schedule (Massed AM n = 25; Massed PM n = 27, Spaced n = 22). This sample size was informed by previous studies examining spaced retrieval practice or word learning in classrooms (e.g., Wegener et al., 2022; Seabrook et al., 2005). The age range was chosen as sleep-based processes that support consolidation (i.e., SWS) have been found to increase around this age (e.g., Ohayon et al., 2004). Thus, we were more likely to observe an overnight change (from Day 1 to Day 2), allowing us to determine whether this was affected by the initial retrieval schedule.

Children in the Massed schedules completed the exposure session and two retrieval practice blocks (B1-2) back-to-back in the morning (10:30) or afternoon (14:00) on Day 1. Children in the Spaced schedule completed the exposure session, followed by one retrieval practice block in the morning (10:30) and then one in the afternoon (14:00) on Day 1. All children then completed a follow-up retrieval practice block on Day 2 (B3), taking place at the same time as the first was on Day 1 (i.e. 10:30 for Massed AM and Spaced, 14:00 for Massed PM). Finally, one week after Day 1, all children completed a final retrieval practice block (B4) at the exact times as the Day 2 blocks were conducted, allowing longer-term measures.

### ***Standardised tests***

At the end of the final retrieval practice block (B4), a set of standardised tests were administered. A definitions task (BAS-3; Swinson, 2013; Elliott, Smith & McCulloch, 1997), was used to measure verbal IQ. Here the experimenter read out age-appropriate words, and the children were asked to write what that word meant in their booklet (e.g., *BED = something you lie in when you sleep*). A matrices task (BAS-3; Swinson, 2013; Elliott, Smith & McCulloch, 1997), was used to measure nonverbal IQ. The children were given a booklet with a number of grids with figures, an empty section, and six potential figures that could fit in the empty space. Their task was to select which of the options should fit in the empty space based on the logic followed in the existing figures. The experimenter provided five examples and ensured that everyone understood the task before completing the remaining trials independently. These tasks were originally intended for individual administration, however, we used an adapted version (from James et al., 2017) which allowed group administration.

The standardised tests were analysed to ensure that potential Schedule effects were not due to differences in verbal and non-verbal IQ. First, based on a group-adapted scoring system (from James et al., 2017), we calculated an average standardised t-score for the three schedules in the definitions and matrices task (see Table 4). Then, the t-scores were used in a Bayesian ANOVA comparing the three schedules (Massed AM, Massed PM, Spaced). The analysis did not support a difference between schedules in verbal IQ scores (definitions $BF_{10}$=0.22, error%=0.03). Similarly, t-scores did not differ between schedules in non-verbal IQ measures (matrices BF10=0.24, error%=0.03). Thus, the children in the three schedules did not differ in verbal or non-verbal IQ, suggesting that potential Schedule effects on retrieval performance were due to the experimental manipulation.

**Table 4.** Mean standardised scores (t-score) in definitions and matrices tasks of Experiment 4.

| **DEFINITIONS**<br>**Schedule** | <br>**Mean t-score (±SD)** |
|---|---|
| Massed AM | 45.8 (±10.9) |
| Massed PM | 44.2 (±10.2) |
| Spaced | 42.4 (±8.9) |
| **MATRICES**<br>**Schedule** | <br>**Mean t-score (±SD)** |
| Massed AM | 53.8 (±13.4) |
| Massed PM | 57.1 (±9.2) |
| Spaced | 52.3 (±12.0) |

#### *3.3.1.2 Material*

The stimuli were names and pictures of real but rare animals intended to be novel to most participants. Nine rare animals were taken from Fletcher et al. (2020), with additional items added to produce a list of twelve new animals. The animal names had already been used with adults (Chapter 2), where no participants reported prior knowledge of them, thus, it was unlikely that children would be familiar with the animal names. Each rare animal was associated with a familiar "base" animal (e.g., e.g. a *TENREC* resembles a hedgehog). See Appendix B1 for the list of animal names and base animals used. The children were presented with the written word (all word presentations were in CAPITAL letters, but children were not encouraged to write in a specific way as we did not consider capitalisation when scoring the responses) and photographic stimuli on a projected computer screen in the classrooms using Microsoft PowerPoint and provided written responses in physical booklets provided by the experimenter.

#### *3.3.1.3 Design*

A mixed design was used: Schedule (Massed AM, Massed PM, Spaced) was a between-subjects variable, and Block (B1, B2, B3, B4) was a within-subjects variable. The dependent variable for each test (cued recall, picture naming, and base animal match) was a matching percentage based on a Levenshtein distance between children's responses and the correct answer (described in more detail in Chapter 2, Section 2.3.1.5).

#### *3.3.1.4 Procedure*

All children completed one initial exposure session and four blocks of retrieval practice on three days. The first day of the experiment (Day 1) started with an initial exposure session in the morning (start time = 10:30) or in the afternoon (start time = 14:00). The exposure session was followed by two Massed or Spaced Blocks of retrieval practice (B1-2). B1-2 took place immediately after the exposure session was finished in the Massed schedules. Children in the Spaced schedule started B1 immediately following the exposure session, after which the experimenter left the classroom to return around 3 hours later for B2 (start time = 14:00). Twenty-four hours later, the experimenter returned for the Day 2 follow-up retrieval practice session (B3; Massed AM and Spaced Day 2 start time = 10:30; Massed PM Day 2 start time = 14:00). Finally, to measure long-term memory performance, the experimenter returned for a final retrieval practice block one week after the first exposure session, on Day 7 (B4; Massed AM and Spaced start time = 10:30, Massed PM start time = 14:00). See Figure 8 for an illustration of the retrieval practice schedules. The exposure and retrieval practice blocks took place in the classrooms of each class, and the usual teacher and teaching assistants were present throughout.

When the experimenter arrived in the classroom, they briefly introduced themselves and the experiment as follows, "*Today we are going to learn the names of new animals, some of these might look a bit like animals you already know. I want you to try your best to learn as many names as possible.*". The exposure session then commenced, with the experimenter reading out the instructions for each task (described later) and ensuring everyone understood the task before presenting the stimuli. After the exposure tasks were finished, the experimenter said, "*Now we are going to see how many of these new animals you can remember. We will do a couple of tests, and in each, you will need to write down the new animal names in your workbook. It might be quite hard to remember all the names, so just make sure you try your best and write down whatever you can remember, even if you are not 100% sure. If you have absolutely no idea of what an animal's name might be, you can have a guess*". The experimenter then read the instructions for the retrieval practice tests and allowed questions before commencing. On Day 2 and Day 7, the experimenter returned and repeated the retrieval practice tests in the third and fourth Block. After the retrieval practice tests on Day 7, the experimenter also conducted standardised verbal and non-verbal IQ measures (definitions and matrices, respectively) to ensure the different classes (and Schedule conditions) were well matched on these ability metrics. After completion, the experimenter provided relevant teachers with small gifts (stickers and erasers) to distribute to the children of the three classes. The teachers were also given an Amazon voucher for classroom equipment as a thank you for participation.

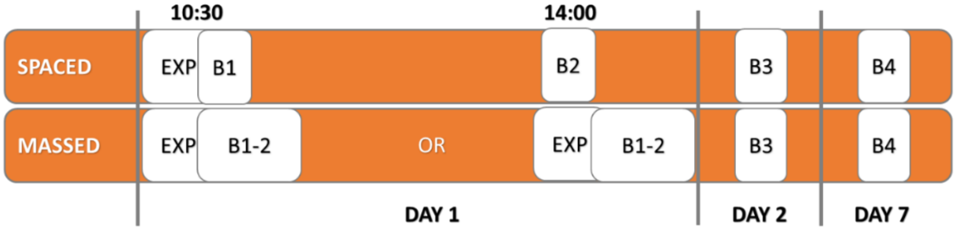


**Figure 8.** Outline of the timetable used with child participants in Spaced and Massed schedules in Experiments 4 and 5. Experiment 5 did not include the Day 2 retrieval block (i.e., B3).

### *Exposure session*

The exposure session (lasting no longer than 15 minutes) consisted of 4 tasks designed to briefly expose the children to the names of the rare animals and their associated images. There were six auditory exposures of the animal name, five exposures to the written version of the animal name, five exposures to the picture, and only three complete exposures (i.e. picture and auditory and written name together).

*Prior knowledge task.* To determine whether the children were familiar with the animal names before the study, they were presented with an audio recording of each animal name and asked to indicate familiarity by circling YES or NO in their booklet. Each animal name was presented once, in a fixed random order, for 12 trials. If a child indicated prior knowledge of an animal, they were asked to write a brief definition (e.g. *ELEPHANT = large animal with big ears and a long trunk*) before the experimenter moved on to the subsequent trial. If they indicated prior knowledge and provided an accurate definition of the animal, the trials of the relevant items were removed from analyses. No participants indicated prior knowledge of the animals, so no items were removed from the analyses.

*Repetition task.* To allow the children to practise writing the animal name, they were presented with an audio recording and the written form of each animal name and asked to copy the name by writing it in their booklet. They were then presented with the name and accompanying picture for 3 seconds before the subsequent trial was initiated. Twelve trials occurred, one for each animal, in a fixed random order.

*Picture selection.* The children were presented with two pictures of the newly learned animals, one target and one distractor, and the written version of the target animal's name. The pictures were labelled A or B, and the children's task was to select which picture belonged to the name by circling the corresponding letter in their booklet. The time limit for making their selection was 15 seconds before being presented with the correct answer for 3 seconds. The target and distractor animals were pre-selected randomly, and it was ensured that no repeated combinations occurred, and the trials were presented in a fixed random order.

*Name selection.* The children were presented with two names of the newly learned animals, one target and one distractor, and one picture of the target animal. Their task was to select which name belonged to the picture by circling the corresponding name in their booklet. The time limit for making their selection was 15 seconds before being presented with the correct answer for 3 seconds. The target and distractor animals were selected randomly, and it was ensured that no repeated combinations occurred, and the trials were presented in random order.

***Retrieval practice blocks***

The retrieval practice blocks contained three tests always administered in the same order: cued recall, picture naming, and base animal match. Each block took 10-15 minutes to complete. Initial exposure was limited to three complete exposures, so the tests were expected to be challenging. Children were encouraged to write down everything they could recall or guess if they were not confident. It was important that at least an apparent attempt at retrieving all animal names was made, as previous literature has shown that brief retrieval attempts can aid later recall (Barcroft, 2007).

*Sleepiness ratings*. Before each retrieval block, the children were asked to indicate how sleepy they were feeling using the Karolinska Sleepiness Scale (Åkerstedt & Gillberg, 1990). The children were presented with a 1-10 scale *(*e.g., *1 = not sleepy at all, extremely alert; 5 = Neither alert nor sleepy; 10 = Extremely sleepy, cannot keep awake*) and asked to circle how sleepy they felt in the 5 minutes before. This measure was included as higher ratings of daytime sleepiness have been found to negatively affect verbal cognitive functions (Macchitella et al., 2020).

*Cued recall test.* This test measured the child's orthographic knowledge of the newly learned animal names. The children were cued with the first two letters of each new animal name (e.g., TE) and asked to retrieve the full name and write it in their booklet (e.g., TENREC). They were given roughly 20 seconds to write their answer before the subsequent trial, but the experimenter could move on faster if everyone had finished writing. The presentation order of the cues was in a fixed-random order, and each cue was presented once.

*Picture naming test.* This test allowed us to measure the association between the picture and new orthographic forms. In random order, the children were presented with a picture of an animal and asked to retrieve and write down its name in their booklet. They were again given 20 seconds to write down their answer, but the experimenter could move on faster if everyone finished before the allowed time limit. Each picture was presented once for 12 trials in total.

*Base animal match test.* This test was inspired by Lindsay and Gaskell (2013) and was intended to encourage the retrieval of the new animal names along with an already known, similar-looking animal to aid learning by integrating novel and existing information through retrieval. The children were presented with an already known animal (e.g. HEDGEHOG) and asked to retrieve and write down the name of a similar-looking animal they learned in the exposure session (e.g. TENREC). After writing their answer, the correct new animal was displayed to provide feedback and limited additional learning. The test consisted of 12 trials, one for each new animal.

#### *3.3.1.5 Calculation of matching percentage (dependent variable)*

The children provided written responses in each retrieval practice test, allowing us to capture more subtle changes in spelling accuracy across tests (as opposed to scoring responses as correct/incorrect). In addition, previous studies have shown that spelling accuracy is sensitive to encoding conditions (e.g., the time between initial learning, recall tests, and sleep; Kurdziel & Spencer, 2016), so this would allow us to capture more discrete changes in the memory performance. We calculated the matching percentage based on the children's responses following the methodology outlined in Section 2.3.1.5.

#### *3.3.1.6 Statistical analyses*

To determine whether performance in the three retrieval tests on Day 1 differed for children in the spaced and massed retrieval practice schedules, we calculated the average matching percentage for each child as outlined above. The matching percentage was used as the dependent variable in a 2(Block: B1, B2) x 3(Schedule: Massed AM, Massed PM, Spaced) Bayesian repeated-measures ANOVA. This analysis was run separately for the three tests (cued recall, picture naming, and base animal matching). Further, additional analyses were conducted to determine whether overnight changes from Day 1 to Day 2 and long-term changes to Day 7 depended on the initial retrieval practice schedule. We ran a 3(Block: B2, B3, B4) x 3(Schedule: Massed AM, Massed PM, Spaced) Bayesian repeated-measures ANOVA for all tests. B2 indicates performance at the end of Day 1, B3 indicates performance on Day 2, and B4 indicates performance one week later, on Day 7.

All analyses were run in JASP version 0.16 (JASP Team, 2022), an open-source software. We maintained the settings for analysis as outlined in Section 2.3.1.6. As per the general guidelines also outlined there, we interpreted a BF of 1 as no support for either hypothesis, BF between 1 and 3 as weak support, BF between 3 and 10 as moderate, and values above ten are considered to be strong support for the alternative hypothesis (van Dorn et al., 2020; Kass & Raftery, 1995; Jeffreys, 1939). Values below 0.3 indicate support for the null hypothesis. We will also report an error% linked to the BF, where lower numbers indicate greater stability in the numerical analysis (values below 20% are generally considered robust, as per van Dorn et al., 2020). All data, analyses and JASP outputs are available for open access at https://osf.io/tvcm9/.

### 3.3.2 Experiment 4 Results

Out of the 74 children participating, only 60 completed all retrieval sessions on the three days. This relatively large change in participation numbers was because the time of the study was close to national lockdowns due to the COVID-19 pandemic, resulting in many children being absent due to isolation regulations. The final number of children in each schedule was: 22 in Massed AM, 19 in Massed PM, and 19 in Spaced. We appreciate that the final number of children in the Schedules was lower than initially expected, increasing the risk of the analyses being underpowered. However, a benefit of using Bayesian statistics is that they can be used reliably on smaller sample sizes (e.g., Aczel et al., 2020; Schmid & Stanton, 2020; Kruschke & Liddell, 2017). We will include BF values for all analyses as the size of the BF support is a good indicator of how powered the analyses are; we will also address the strength of these in a later section. Furthermore, Experiment 5 (which took place 1.5 years later, after many national restrictions were lifted) was sufficiently powered, allowing us to retest the key hypotheses tested in Experiment 4.

The participating children were asked not to change their answers after receiving feedback (i.e., the correct answer) in the base animal match test. However, and unfortunately, there were clear indications that some initial answers had been changed, or the child had waited to write their answer until feedback had been presented. For example, there were instances where a child could recall very few (e.g. <2) animal names correctly in the picture naming test, but immediately after, in the base animal match test, they could recall 100% of the names. This improvement was not likely due to the test order as we did not see this pattern with an adult population (Chapter 2) using the same design but without the opportunity to change their answer after the feedback presentation. We could not completely determine which children had changed their answers, so we decided not to include the base animal match test in further analyses. For completion, these data can be found at https://osf.io/tvcm9/.

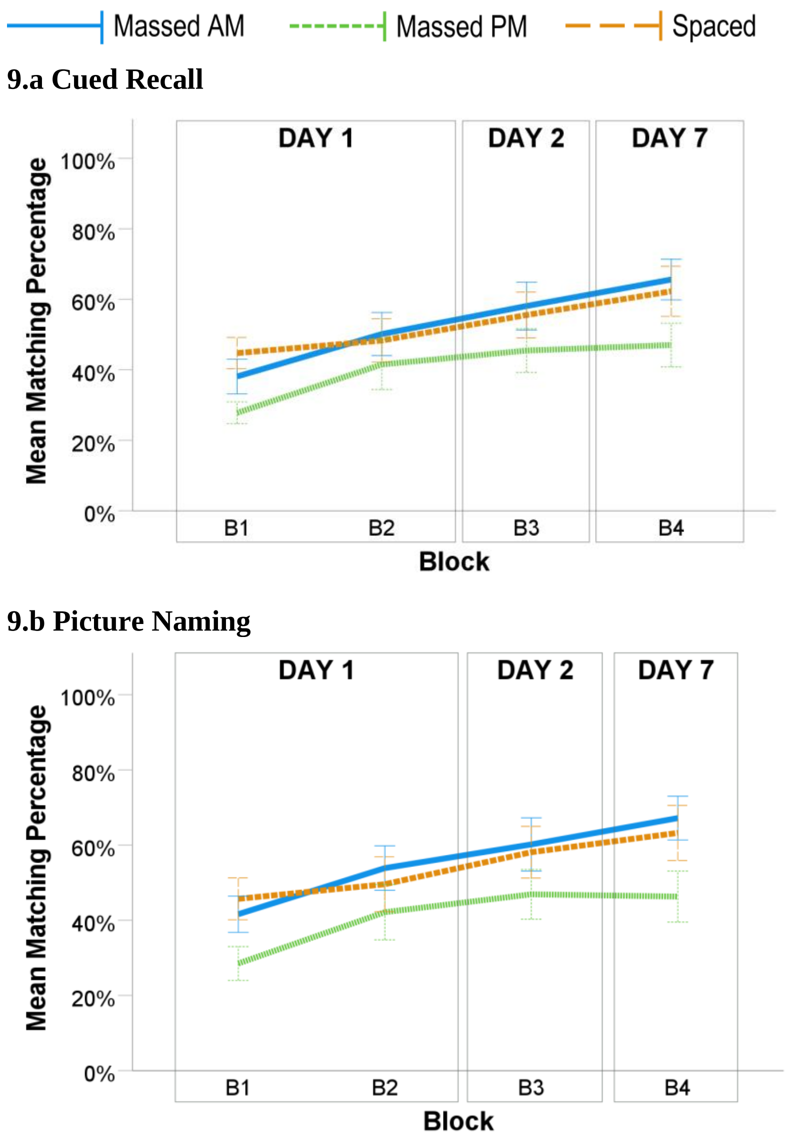
#### *3.3.2.1 Whether within-day spaced retrieval practice aid learning and elicited a spacing benefit on Day 1 (B1-2)*

The first analysis determined if overall performance improved from B1 to 2 and if there was a spaced benefit on Day 1. A 2(Block 1, 2)x3(Schedule) Bayesian repeated-measures ANOVA was conducted with the average matching percentage as the dependent variable. See Table 5 for the mean matching percentage (± SD) for Experiments 4-5. The analyses revealed overwhelming support for changes in performance across Block in both the cued recall and picture naming tests (cued recall $BF_{10}$=81.89, error%=1.54; picture naming $BF_{10}$=1872.8, error%=0.97). However, there was no support for an effect of Schedule (cued recall $BF_{10}$=0.44, error%=0.55; picture naming $BF_{10}$=0.65, error%=2.86). Additionally, there was no support for the Block*Schedule interaction in either test (cued recall $BF_{incl}$=0.59; picture naming $BF_{incl}$=0.54), indicating that children showed a similar change in performance (i.e., improvements) regardless of which schedule they were assigned to (as seen in Figure 9).

#### *3.3.2.2 Whether within-day spaced retrieval practice results in a spacing benefit one day and one week after initial exposure (B2-4)*

Our second set of analyses aimed to determine whether performance changed depending on the retrieval practice schedule when measured 24 hours (i.e., Day 2) and one week (i.e., Day 7) later. Again, the average matching percentage was used as the dependent variable in a 3(Block 2, 3, 4)x3(Schedule) Bayesian repeated-measures ANOVA. The analysis revealed overwhelming support for an effect of Block, suggesting changes from the end of Day 1 to Day 2 and Day 7 in the retrieval tests (cued recall $BF_{10}$=40400.1, error%=0.64; picture naming $BF_{10}$=348416.9, error%=0.95). Post-hoc tests were conducted on the Block variable to explore this effect further. The post-hoc analyses provide an unadjusted BF; however, to correct for multiple comparisons, the posterior odds (PO) were reported as they corrected for multiple testing by setting the prior odds (i.e., the probability of no effect) to 0.5 (van den Bergh et al., 2019; Westfall, Johnson & Utts, 1997). Post-hoc analyses revealed strong to overwhelming support for changes from Day 1 to Day 2 in all tests (cued recall PO=6.05, error%=4.3e-7; picture naming PO=31.92, error%=7.1e-9), from Day 1 to Day 7 (cued recall PO=1084.9, error%=5.2e-10; picture naming PO=13887.77, error%=1.99e-10), and from Day 2 to Day 7 (cued recall PO=36.84, error%=6.9e-8; picture naming PO=7.30, error%=1.69e-7). Figure 9 shows that children in all schedules could recall more words across the days (i.e. continuous improvements), resulting in the best performance on Day 7.

In contrast to our hypotheses regarding a spaced benefit on Day 2 and 7, there was weak support for the null hypothesis in the Schedule variable in either test (cued recall $BF_{10}$=0.52, error%=0.66; picture naming $BF_{10}$=0.68, error%=12.41), and strong support for the null hypothesis in the Day*Schedule interaction in the Cued Recall test ($BF_{incl}$=0.10), and weak support in the Picture Naming test ($BF_{incl}$=0.39). Thus, the retrieval practice Schedule did not influence long-term measures one week later.



**9.a Cued Recall**

**9.b Picture Naming**

**Figure 9. Experiment 4 results.** Performance changes in mean matching percentage for participants in the spaced and massed schedules in the cued recall (**9.a**) and picture naming (**9.b**) tests across the four blocks on Days 1-7. Error bars indicate ±1SE.

It should be noted that our concern about power was somewhat confirmed by low BFs, indicating that the null hypothesis was only between 1.5-1.7 times more likely to occur than the alternative hypothesis for the effect of Schedule. While this still supports the null hypothesis, it indicates that the sample size in each Schedule was most likely on the low end.

**Table 5. Mean matching percentage (±SD) of Experiments 4 and 5.** Table shows mean matching percentage across retrieval Blocks (B1-2 occurred on Day 1, B3 on Day 2, and B4 on Day 7) for participants in the different Schedules (Massed AM, Massed PM, Spaced) in Experiments 4 and 5.

| | DAY 1 | | DAY 2 | DAY 7 |
|---|---|---|---|---|
| | **B1** | **B2** | **B3** | **B4** |
| **CUED RECALL** | | | | |
| **Experiment 4** | | | | |
| Massed AM | 38.1 (±22.6) | 50.2 (±28.0) | 58.1 (±31.1) | 65.6 (±26.4) |
| Massed PM | 27.8 (±13.5) | 41.5 (±30.9) | 45.5 (±27.0) | 47.0 (±27.0) |
| Spaced | 44.7 (±19.4) | 48.3 (±26.9) | 55.6 (±28.4) | 62.3 (±31.0) |
| **Experiment 5** | | | | |
| Massed | 31.0 (±18.4) | 47.2 (±23.6) | | 48.1 (±23.6) |
| Spaced | 30.3 (±15.6) | 34.8 (±14.7) | | 45.2 (±16.2) |
| | **DAY 1** | | **DAY 2** | **DAY 7** |
| | **B1** | **B2** | **B3** | **B4** |
| **PICTURE NAMING** | | | | |
| **Experiment 4** | | | | |
| Massed AM | 41.6 (±22.6) | 53.9 (±27.7) | 60.2 (±33.2) | 67.2 (±27.3) |
| Massed PM | 28.5 (±19.6) | 42.2 (±32.1) | 46.9 (±28.8) | 46.3 (±29.6) |
| Spaced | 45.7 (±24.3) | 49.6 (±31.9) | 58.1 (±30.0) | 63.2 (±31.9) |
| **Experiment 5** | | | | |
| Massed | 34.3 (±18.1) | 49.3 (±25.4) | | 51.6 (±24.5) |
| Spaced | 33.8 (±16.1) | 40.0 (±18.5) | | 46.7 (±18.9) |

### *3.3.2.3 Sleepiness ratings*

We analysed the KSS ratings that were provided by each child before starting each retrieval Block, in a 4(Block) x 3(Schedule) Bayesian repeated-measures ANOVA. Of central interest was the potential effect of Schedule, as this would indicate that the children reported different levels of sleepiness in the different schedules. The analysis supported an effect of Schedule ($BF_{10}$=396.7, error%=1.02), with posthoc tests indicating that both Massed schedules reported higher sleepiness ratings than those in Spaced (see Table 6), with Massed PM reporting the highest levels of sleepiness. As we did not observe an effect of Schedule in the previous analyses, the differences in sleepiness ratings were likely not enough to drive statistically significant differences in performance between Schedules.

**Table 6. Mean KSS ratings (±SD) of Experiments 4 and 5 and Schedule post-hoc tests for Experiment 4.** Table shows sleepiness ratings in retrieval Blocks (B1-2 occurred on Day 1, B3 on Day 2, and B4 on Day 7) for participants in the different Schedules (Massed AM, Massed PM, Spaced) of Experiments 4 and 5. The results of the post-hoc tests in Experiment 4 are included at the bottom.

| | **DAY 1** | | **DAY 2** | **DAY 7** |
|---|---|---|---|---|
| | **B1** | **B2** | **B3** | **B4** |
| **Experiment 4** | | | | |
| Massed AM | 4.0 (±2.3) | 5.1 (±2.9) | 4.5 (±2.6) | 3.5 (±2.6) |
| Massed PM | 6.1 (±3.3) | 7.4 (±2.9) | 4.6 (±3.2) | 6.2 (±3.3) |
| Spaced | 2.5 (±1.9) | 3.4 (±2.8) | 3.3 (±2.4) | 2.9 (±2.2) |
| **Experiment 5** | | | | |
| Massed | 5.2 (±2.4) | 4.9 (±3.1) | | 5.1 (±3.1) |
| Spaced | 5.4 (±3.0) | 6.2 (±3.3) | | 5.3 (±3.1) |

| | | **Unadjusted BF$_{10}$** | **PO** | **Error%** |
|---|---|---|---|---|
| **Post-hoc tests, Experiment 4** | | | | |
| Massed AM | Vs Massed PM | 114.7 | 67.37 | 1.9e-9 |
| | Vs Spaced | 16.94 | 9.95 | 1.5e-8 |
| Massed PM | Vs Spaced | 1.4e+7 | 8.1e+6 | 5.0e-10 |

### *3.3.2.3 Exploratory analysis of verbal IQ and overnight change in word recall*

Previous studies have found that expressive vocabulary knowledge (as measured in the definitions task in the current experiment) can correlate with overnight consolidation effects (e.g., Henderson et al., 2015) and accurately predict improvements in word learning tasks after sleep (James et al., 2017). Thus, children with better vocabulary knowledge typically also show bigger overnight gains from consolidation processes. To explore this here, we collapsed performance across the three schedules conditions and conducted a Bayesian Kendall Tau correlation of the overnight change (*overnight change = performance at B3 – performance at B2*) and the standardised t-score in the definition task. Contrary to previous findings, our analysis did not reveal support for a correlation (cued recall tau=-0.03, BF$_{10}$=0.20; picture naming tau=0.14, $BF_{10}$=0.47), indicating that individual differences in vocabulary did not correlate with benefits from overnight consolidation here. There could be several reasons for this discrepancy. For example, studies that previously found a correlation used a different type of word-learning task (story-book learning; Henderson et al., 2015; James et al., 2017) and measure of word learning (i.e., lexical competition, Henderson et al., 2015). Speculatively, vocabulary might be more strongly associated with consolidation when the language demands at the point of learning are richer (i.e., when learning from story contexts, e.g., Henderson & James, 2018) and/or when more sensitive measures of lexical integration capture word learning changes.

### 3.3.3 Experiment 4 Discussion

The aim of Experiment 4 was to determine whether children show a spacing benefit from within-day spaced retrieval practice without intervening sleep and, crucially, whether this benefit is also observable one day and one week later. Contradicting our hypotheses, the results showed no indication of a spacing benefit at any test point. However, due to the relatively low BF values, we cannot with certainty fully accept the null hypothesis. Instead, we found improvements across all retrieval practice blocks in the cued recall and picture naming tests regardless of the retrieval schedule on Day 1. The finding that children show improvements over repeated tests is consistent with other developmental literature suggesting that children continue to improve their memory for new words over time and sleep (e.g. James et al., 2020; Williams & Horst, 2014).

Surprisingly, we did not observe either a spaced or massed benefit on Day 1, as we have observed both in extant literature (e.g. Childers & Tomasello, 2002, reported an initial spacing benefit; while Petersen-Brown et al., 2019, reported an initial massed benefit). This finding does not follow the available theories on spaced retrieval practice in either adults or children. Additionally, as outlined in Chapter 2, we observed an initial steeper learning trajectory for adults practising retrieval in massed blocks. In the current experiment, we observed a similar trend toward a steeper learning trajectory in the Massed schedules, but the statistical analyses did not support this. A potential explanation for the lack of statistical support in Experiment 4 could be due to the study being underpowered (as indicated by the low support for the null hypothesis). This issue will be addressed in Experiment 5. It could also potentially be due to the children in the Massed condition reporting higher levels of daytime sleepiness than children in the spaced schedule (as seen in the analysis of the KSS task). Higher daytime sleepiness has been related to lower vocabulary cognitive function in school-aged children (Macchitella et al., 2020), thus, potentially explaining why children in the massed schedule did not improve as much as they could have if their sleepiness levels were at an equal level as the children in the spaced schedule.

Surprisingly, we did not observe a spacing benefit one week after initial exposure, contrasting other classroom-based studies (e.g. Goossens et al., 2014; Sobel et al., 2011). However, one possibility is that the follow-up session on Day 2 allowed for further retrieval practice, providing a spaced retrieval opportunity to all participants (i.e. essentially making all groups spaced over two consecutive days). We applied this same logic as in Experiments 2 and 3 in Chapter 2. These experiments showed that including a retrieval opportunity on Day 2 significantly impacted the trajectory of performance from Day 1 to Day 7. While we did not observe a spacing benefit in the adult population, we found strong support that within-day spaced retrieval practice protected memories from forgetting when no additional retrieval opportunity was provided after sleep. In contrast, an additional retrieval opportunity on Day 2 was needed to prevent forgetting if initial retrieval practice was massed. Thus, there is strong support that including a retrieval practice block on Day 2 (i.e., after a period of sleep) affected later retention levels one week later. Since sleep-based consolidation effects (and potentially reconsolidation) can have a greater effect on memory retention in children (James, Gaskell & Henderson, 2019; James et al., 2017; Wilhelm et al., 2012), it is important to re-examine whether this was the case in children as well, by removing the Day 2 retrieval block.

## 3.4 Experiment 5

Experiment 5 built on Experiment 4 by ensuring it was sufficiently powered and exploring whether school-aged children show long-term benefits in recall performance from within-day spaced retrieval practice of new words when no post-sleep retrieval opportunity was given.

Based on Experiment 4, we hypothesised that Spaced and Massed schedules would improve from Block 1 to Block 2 on Day 1. We also expected a potential interaction of Block*Schedule but in the reverse direction of our prediction in Experiment 4 (i.e., we predicted here that there would be a massed benefit by Day 1 due to a steeper learning trajectory). As discussed above, there was an indication of a trend towards steeper improvement for the Massed schedules than in Spaced, however, this did not receive statistical support, potentially due to Experiment 4 being underpowered. In Experiment 5, we ensured that the final sample size was more similar to prior studies, increasing the statistical power. We also observed this pattern in adults using the same study design (Chapter 2); thus, we hypothesise that a potential benefit for the Massed schedule would emerge by Day 1 (i.e., children in the massed schedule would perform higher than spaced in B2). Finally, without the Day 2 retrieval practice opportunity, we hypothesised that a spacing benefit would emerge one week later, manifesting as better performance in the cued recall and picture naming tests due to continuous improvements or less forgetting than from massed retrieval practice (confirmed by a Day*Schedule interaction).

### 3.4.1 Experiment 5 Methods

The methods were identical to Experiment 4, except for removing the follow-up Block on Day 2 (i.e. B3). In addition, we combined the Massed AM and PM schedules to a single Massed group (due to them showing the same pattern of change in Experiment 4) in an effort to reduce interruptions to the normal school-day post-Covid. The methods and analyses of Experiment 5 were preregistered at https://osf.io/s4azr.

#### *3.4.1.1 Participants and material*

Fifty-six children in years 4 and 5 at a Yorkshire (UK) primary school participated in the study (mean age = 9 years, 9 months). Half the children in each class were assigned to the Spaced schedule (n=30), and the remaining halves to Massed (n=26). All children completed the required retrieval blocks, resulting in 56 complete datasets. We used the same material and tasks as Experiment 4.

#### *Standardised tests*

To ensure that the children in the different schedules were matched on verbal and non-verbal IQ, we conducted and analysed the standardised tests in the same way as in Experiment 4 (the average t-scores can be seen in Table 7). The analysis of the definition task strongly supported the null hypothesis ($BF_{10}$=0.27, error%=0.01), indicating that the Schedules did not differ in verbal IQ. Similarly, there was strong support for the null hypothesis in the matrices task ($BF_{10}$=0.30, error%=0.01). Thus, the potential effects in the later analyses were not likely due to individual differences in verbal and non-verbal IQ.

**Table 7.** Mean standardised scores (t-score) in definitions and matrices tasks of Experiment 5.

| **DEFINITIONS**<br>**Schedule** | **Mean t-score (±SD)** |
|---|---|
| Massed | 46.9 (±11.0) |
| Spaced | 47.3 (±12.1) |

| **MATRICES**<br>**Schedule** | **Mean t-score (±SD)** |
|---|---|
| Massed | 58.4 (±10.4) |
| Spaced | 59.7 (±9.2) |

#### *3.4.1.2 Design and procedure*

The design and procedure was the same as experiment 4, except the Day 2 retrieval practice block (B3) was removed.

#### *3.4.1.3 Statistical analyses*

The matching percentage was used as the dependent variable in all analyses and calculated in the same way described in Experiment 4. To test changes in performance across the two retrieval sessions on Day 1, we ran a 2(Block) x 2(Schedule) Bayesian repeated-measures ANOVA. To measure long-term changes in performance from Day 1 to Day 7, we conducted a 2(Block 2-4) x 2(Schedule) Bayesian repeated-measures ANOVA on the three tests.

### 3.4.2 Experiment 5 Results

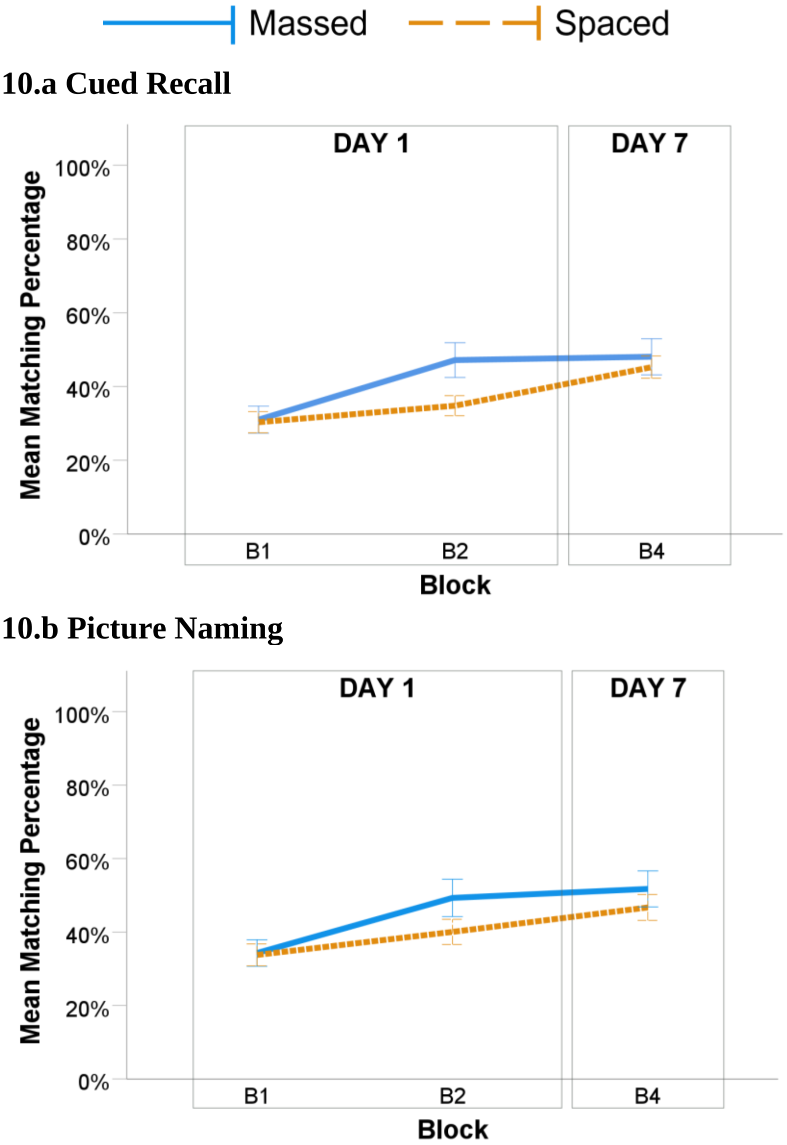
Two outliers were identified (one from Spaced and one from Massed) in the cued recall and picture naming measures, and the datasets were removed from the relevant analyses.

#### *3.4.2.1 Whether repeated retrieval practice aids learning and elicits a massed benefit on Day 1 (B1-2)*

The first analysis investigated if overall performance improved from B1 to 2 and if there was an initial massed benefit prior to sleep. A 2(Block 1-2)x2(Schedule) Bayesian repeated-measures ANOVA was conducted with the average matching percentage as the dependent variable. The analyses revealed overwhelming support for changes in performance across Blocks in both the Cued Recall and Picture Naming tests (cued recall $BF_{10}$=2513.7, error%=2.79; picture naming $BF_{10}$=1994.8, error%=1.14). However, there was small support for the null hypothesis in the Schedule variable (cued recall $BF_{10}$=0.73, error%=0.92; picture naming $BF_{10}$=0.51, error%=1.79). In addition, the analysis revealed strong support for the Block*Schedule interaction in the cued recall test ($BF_{incl}$=13.94) but no support in the picture naming test ($BF_{incl}$=1.65). Figure 10 shows that all participants improved their performance from B1 to B2, but those in the Massed schedule improved at a steeper rate than those in the Spaced schedule for the cued recall test.

#### *3.4.2.2 Does within-day spaced retrieval practice result in a spacing benefit one week after initial learning when no post-sleep retrieval opportunity was given? (B2-4)*

The average matching percentage was used as the dependent variable in a 2(Block2,4)x2(Schedule) Bayesian repeated-measures ANOVA. The analyses revealed strong support for changes in performance across the Blocks for the cued recall test ($BF_{10}$=61.29, error%=1.06) and medium support for picture naming ($BF_{10}$=3.83, error%=1.32). There was little support for the null hypothesis in Schedule (cued recall $BF_{10}$=0.89, error%=2.44; picture naming $BF_{10}$=0.76, error%=4.70). To determine whether the children changed performance at different rates from Day 1 to Day 7 depending on which retrieval practice schedule they were assigned to, we focused on the Block*Schedule interaction. The analysis revealed strong support for the interaction in the cued recall test ($BF_{incl}$=17.59) but no support in the picture naming test ($BF_{incl}$=0.49). Figure 10 shows that children in the Spaced schedule improved their performance from Day 1 to Day 7 while children in the Massed schedule maintained performance in the cued recall test. However, and importantly, the Spaced schedule did not perform higher than Massed; they merely caught up to the same level of performance as Massed had reached on Day 1.



10.a Cued Recall

10.b Picture Naming

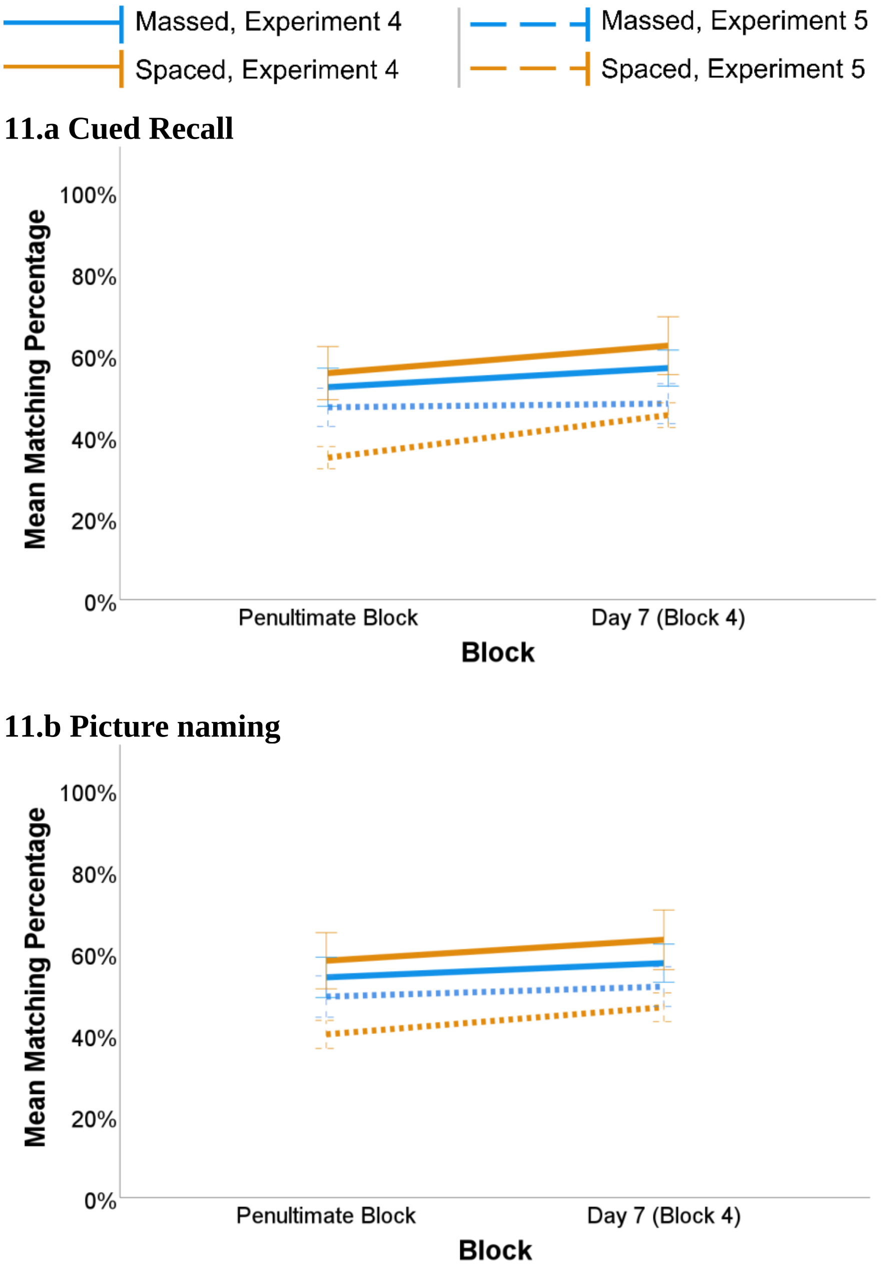
**Figure 10. Experiment 5 results.** Mean matching percentage of children in the Spaced and Massed schedule in the three blocks across Day 1 and 7, in the cued recall (**10.a**) and picture naming (**10.b**) tests.

#### *3.4.2.3 Sleepiness ratings*

We analysed the KSS ratings in a 3(Block) x 3(Schedule) Bayesian repeated-measures ANOVA, see Table 6 for the mean KSS ratings. There was weak support for the null hypothesis in the Schedule variable ($BF_{10}$=0.43, error%=0.63), suggesting that daytime sleepiness ratings did not differ between the schedules. Thus, the effects above were unlikely to be linked to differences in sleepiness.

#### *3.4.2.4 Exploratory cross experiment comparison: Experiments 4 and 5*

To examine whether a post-sleep retrieval opportunity after within-day spaced or massed retrieval practice affects long-term retention, we explored whether performance changed from the penultimate Block on Day 1 (i.e. B3 in Experiment 4, and B2 in Experiment 5) to the final Block on Day 7 (results can be seen in Figure 11). We ran a 2(Experiment: Experiment 4, Experiment 5) x 2 (Schedule: Massed, Spaced) x 2 (Block: Penultimate B, Final B) to determine whether long-term changes in performance was affected by both initial retrieval schedule and whether or not a post-sleep retrieval opportunity was provided on Day 2. There was no support for an interaction effect of Experiment *Schedule*Block in either the cued recall or picture naming task (cued recall $BF_{incl}$=1.00; picture naming $BF_{incl}$=0.35), suggesting that Day 7 performance was not affected by an extra retrieval opportunity after sleep and the initial retrieval practice schedule.



11.a Cued Recall

11.b Picture naming

**Figure 11. Results of Experiment 4 and 5 comparison analysis.** Data shows changes in mean matching percentage performance in cued recall (**11.a**) and picture naming (**11.b**) from the penultimate retrieval block (i.e. B3 in Experiment 4 and B2 in Experiment 5) to the final retrieval block (B4) on Day 7. Solid lines represent Experiment 4; dashed lines represent Experiment 5. Blue lines represent Massed Schedules; orange lines represent Spaced Schedules. Error bars represent ± 1SE.

### 3.4.3 Experiment 5 Discussion

Experiment 5 revealed that spacing two retrieval practice opportunities within a single day does not lead to significantly better word-learning performance than massed retrieval practice in school-aged children, either by the end of that day or one week later. Confirming our hypothesis, we observed an initial improvement in performance from retrieval practice (regardless of schedule) and an initial benefit from massed retrieval practice. Interestingly, however, and in contrast to our hypothesis, we did not observe a performance benefit from spaced (than compared to massed) retrieval practice at the one-week test.

However, we did find support for an interaction, such that children in the Spaced schedule showed greater improvements from Day 1 to Day 7 whereas children in the Massed schedule maintained their performance without improvements. Importantly, this was observed despite no evidence of the massed group performing close to ceiling levels. This suggests that the initial retrieval practice schedule impacted subsequent *changes* in performance, in the absence of exerting an overall performance benefit on Day 7. Specifically, spacing retrieval practice over a day led to a steeper trajectory of performance improvement after Day 1, but only to the extent that they caught up to the massed group one week later.

Interestingly, and quite surprisingly, the cross-experiment comparison did not suggest that an additional Day 2 retrieval practice opportunity influenced performance on Day 7. Thus, it appears that performance one week after initial exposure was not significantly affected by the presence/absence of an additional retrieval opportunity on Day 2. It is possible that the repeated tests on Day 1, in combination with potential offline consolidation processes after Day 1 were sufficient to maintain later performance (Henderson, Weighall & Gaskell, 2013).

### 3.5 General discussion

In two experiments, we investigated whether within-day spaced retrieval practice is more beneficial than massed retrieval practice when children are learning new words and their meanings. Experiment 4 observed improvements over repeat tests but we did not find any differences between spaced versus massed retrieval practice at any test point, contrary to our predictions. A potential reason for the lack of spacing benefits could be the follow-up session on Day 2, which allowed potential consolidation and reconsolidation processes to enhance performance for participants in both the spaced and massed schedules. Therefore, we removed the Day 2 session in Experiment 5 to better isolate the potential effects of within-day spaced (or massed) retrieval practice. In Experiment 5, we observed performance differences between the schedules on Day 1, manifesting as a steeper improvement rate for participants in the massed schedule. This was expected as we had observed an indication of a massing benefit in Experiment 4, but this did not receive statistical support. Focusing on changes from Day 1 to Day 7, we still did not find a traditional spacing benefit (i.e. higher performance after spaced retrieval practice at the one week test). However, there was support for an interaction, where children in the Massed schedule maintained their already high performance, while children in the Spaced schedule showed continuous improvements, closing the performance gap by Day 7. Thus, despite revealing that spaced versus massed scheduling influenced the trajectory of change in memory for new words over seven days, the findings nevertheless show that this influence was not enough to promote better memory retention following spaced scheduling one week later.

### 3.5.1 Theoretical implications

The existing theories outlined in the introduction are largely grounded in data on spaced retrieval practice in adults and laboratory studies. Therefore, it is unclear whether they can be applied to developmental populations and classroom-based studies.  The lack of a reliable spacing benefit in the current experiments indicates that theories behind the spacing effect are not robust enough to counter potential confounds that come with group testing. Further studies are required to determine when a spacing benefit emerges and when it does not emerge in group settings, especially classrooms.

There was support for an initial benefit of massed retrieval practice in Experiment 5, which requires further replication. The retrieval effort and contextual variability theory can explain why children in the massed schedule improved their performance at a steeper rate than those in the spaced schedule). According to the retrieval effort theories, retrieval would be easier in massed conditions as less time has passed between retrieval attempts, allowing short-term memory stores to aid initial retrieval (Pyc & Rawson, 2009). The contextual variability theories suggest that more contextual cues would be available to aid retrieval in the massed conditions, resulting in initially better performance than spaced (Maddox, 2016). While this pattern fits with findings of adult studies (e.g. Miles et al., 2014), and to a certain extent to the data here (i.e., for Experiment 5 but not convincingly for Experiment 4), it contrasts with previous studies that have reported an early emerging spacing benefit in children. For example, Childers and Tomasello reported that children learning novel words in spaced sessions on three consecutive days could recall the words better at the end of the test sessions than if learned in a single massed session on a single day. However, as mentioned in the introduction, this and many other studies reporting an initial spacing benefit include at least one sleep period between the spaced sessions. Thus, the lack of consolidation opportunities between our retrieval blocks on Day 1 could potentially explain why we did not observe the same pattern on Day 1, like other child studies have reported.

An alternative explanation for the steeper improvements for children in the massed schedule could be that the learning load was manageable for the children here relative to previous studies. For example, it has been argued that the early spaced retrieval practice benefits seen in other studies because the learning load has been too large for children to learn many words in a single massed session (Smith & Scarf, 2017). Thus, breaking learning up over several sessions (i.e. spacing retrieval practice blocks) allows children to learn the new material more effectively, resulting in an earlier benefit from spaced practice than massed. Instead, in the current experiments, 12 animal names may have been manageable for the children to learn in a single session, resulting in better performance on Day 1 for children in the massed schedules. Additionally, children can use prior knowledge (i.e. existing words) to link novel and existing words, aiding learning (James et al., 2018). In the current experiment, children could use their prior knowledge of words and their prior knowledge of the base animals linked to the novel animal names. Thus, we provided two opportunities to link the novel animal names with existing knowledge, reducing the need for the time between learning occasions and potentially explaining why the children could learn the animal names more effectively when in the massed schedule. This explanation remains to be tested directly in future studies that manipulate the number of stimuli to be learned.

Despite the pattern of change on Day 1 being explained by the theories of spaced retrieval practice, we did not observe a spacing benefit in the measures one week later. Instead, in Experiment 4, we observed continued improvements for participants practising retrieval in both spaced and massed schedules. This pattern of continued change aligns with findings of improvements after sleep periods in children (e.g. James et al., 2019; Williams & Horst, 2014) and indirectly supports Smith and Scarf's argument that consolidation and reconsolidation could aid retention after spaced retrieval practice. Further, in Experiment 5, after successfully isolating the long-term effects of the initial retrieval schedule, we still did not observe a spacing benefit (i.e. higher performance after practising retrieval in a spaced manner). While there were continued improvements from Day 1 to Day 7 after spaced retrieval practice, these were insufficient to elicit a traditional spacing benefit. Thus, the contextual variability and retrieval effort theories of the spacing benefit cannot explain why we did not observe a spacing benefit in our experiments. This enhances our previous argument that the spacing benefit is a complex phenomenon, requiring continued research to determine better a theoretical understanding of when a spacing effect emerges and when it does not.

Further, the discrepancy between our lack of spacing benefit and the extant literature reporting a spacing benefit after retrieval practice spaced over sleep periods provide indirect support for Smith and Scarf's (2017) argument that consolidation/reconsolidation is an important aspect of explaining the spacing effect. This highlights the importance of continued research of within-day spaced retrieval practice to determine when a spacing benefit emerges and when it does not and for the literature to clearly distinguish between sleep and no-sleep spaced retrieval practice.

In summary, the current experiments shed doubt on the existing theories as they cannot explain why we did not see a spacing benefit in our classroom-based experiments with children. This indicates that the spacing benefit is a nuanced and complex phenomenon, with boundary conditions that are in need of systematic examination.

### 3.5.2 Educational implications for children

Spaced retrieval practice has been promoted as an ideal educational method, but more research is needed before we can endorse school implementation. Whilst our two experiments demonstrate that spaced retrieval practice schedules can be conducted in a group-based school setting and implemented within the school day, the current findings did not specifically favour the use of spaced retrieval practice in educational settings. We did not observe a spacing benefit after within-day spaced retrieval practice; instead, we saw equal performance after one week for both spaced and massed retrieval practice. However, it is possible that performance would continue to improve after spaced retrieval practice, eventually resulting in higher performance in a final retention measure after a more extended period (e.g. one month). Currently, this is purely speculative as we did not include further long-term retention tests but form an important direction for future research to further our understanding of when spaced retrieval practice is beneficial for long-term memory retention and when it is not.

#### *3.5.2.1 Classroom studies*

There has been an increase in classroom studies exploring the spacing effect in school-aged children in the last couple of decades, most of which report a spacing benefit (e.g., Goossens et al., 2012; Goossens et al., 2014; Sobel et al., 2011). However, some classroom-based studies have not reported a traditional spacing benefit (e.g., Goossens et al., 2016; Toppino & Cohen, 2009). For example, Goossens et al. (2016) conducted a classroom study with children in grades 2, 3, 4 and 6. Here they examined distributed restudy and retrieval practice in two sessions across days spanning one week (short lag condition) or four sessions across two weeks (long lag condition) and measured performance one and 11 weeks later. One week later, they found a benefit for retrieval practice in all but one grade and a benefit in the short lag condition in two grades. Interestingly, they did not find a benefit for either condition in the 11-week test and no difference for restudy or retrieval practice. Similar to our findings in Chapter 3, their findings contrast with other literature's failure to find their expected spacing benefit. However, a significant limitation of their study is that all their conditions spanned a sleep period, so did not provide an accurate comparison of spaced and massed retrieval practice. Further, this shows that once sleep is included between sessions, long-term performance can be boosted in classroom settings, regardless of one or multiple nights' sleep occur. This can be likened to Experiments 3 and 5, where we show that spaced or massed retrieval practice does not result in different performance one week later when no sleep occurred between retrieval attempts.

Support for a benefit in memory retention from spaced retrieval practice can be found in classroom-based studies in adults (e.g., see Hopkins et al., 2016 for a spacing benefit in adults' mathematics knowledge), but contrasting findings have also been reported. For example, Grieving and Richter (2018) found a long-term benefit in the retention of lecture material when groups of university students were tested using short-answer tasks but not when from multiple-choice tests. Thus, when tested in a group setting, adults may not benefit from all types of tests, even when an effect emerges in laboratory settings (e.g., Adesope, Trevisan & Sundararajan, 2017).

#### *3.5.2.2 Limitations and future directions of classroom testing*

The above discussion on classroom studies raises the question: why do some classroom-based studies report a spacing benefit but others do not? As we have mentioned, this could be due to differences in procedural measures, such as spacing retrieval practice across days that include intervening sleep. However, it could also be due to the nature of classroom testing versus laboratory testing.

Classroom testing has high ecological validity, but as it is a realistic setting, some aspects could negatively impact the chances of finding effects. One such aspect is to do with less experimental control over the timing spent on each task and trial. While we had a max time (30 seconds) for the retrieval practice tests, some trials were finished quicker than others. When too much time passed between trials, there was a tendency for the children to get impatient and start talking. To avoid the children talking too much and potentially sharing their answers due to too much time between trials, the experimenter decided to move through the trials as soon as all children indicated that they had finished writing. This meant that some trials stayed on the screen for longer than other trials, depending on how fast they were at writing their answers.

Further, as mentioned above, the classroom setting allowed less experimental control over noise and distractions during the tests. In our experiments, this was mainly an issue in the base animal match task. As previously outlined, we did not use the base animal match analyses as we could not control if participants changed their responses after being presented with the feedback. It was also difficult for the children to stay quiet during this test, as once the feedback was provided, many cheered if they remembered it correctly, and some expressed annoyance if they could not remember the animal name correctly. While this provided a cheerful atmosphere in the classroom, it was lively and, in some cases, loud. This could have affected the children's concentration for the subsequent trials and potentially their focus on the other tests. This would not have been an issue if the experiments had occurred individually or in a laboratory. It is possible that the lack of concentration due to distractions from classmates in the classrooms could explain why we did not see a spacing benefit in Experiment 5. Spaced retrieval attempts were assumed to be harder than in the massed schedule (Cepeda et al., 2006), thus, more attention would have been required in each retrieval attempt. If attention was lower, less effort could have been applied in the spaced retrieval attempts, explaining our lack of spacing benefit one week later.

Additionally, Experiment 4 took place in the immediate days before all schools in the UK closed due to the Covid-19 pandemic, and there were marked levels of increased uncertainty and anxiety in the classrooms. This was especially the case in the Massed PM condition, where the atmosphere in the classroom was markedly livelier, and the experimenter noted an increased level of anxiety in that class compared to the other. In addition, many children were absent in the later sessions due to testing positive for covid, resulting in a lower-than-expected sample number. Finally, while we did not have a measure for overall anxiety in the classroom, we found higher levels of daytime sleepiness in children in the Massed PM schedule, and it is possible that this affected how much attention was dedicated to the tests (further supported by the overall lower performance levels compared to children in the other schedules).

Goossens et al. (2016) report similar limitations with their study, further highlighting the issues with applying laboratory-developed methods in educational settings and expecting the same effects to emerge. Thus, we need further studies that examine educationally relevant measures in real-life settings to endure that the theoretical accounts can account for effects in less controlled settings. Future studies could consider including a measure of attention in their procedure, for example, a psychomotor vigilance task (PVT). Ideally, this would have to be adapted to be administered in a group setting (similar to how we used an adapted version of standardised tests in Experiments 3 and 5) or be conducted individually but still in the classroom.

For example, each child could complete a PVT task on technology such as iPads, which are now frequently used in schools as part of their educational plan. Further, we need to conduct similar group-based studies with adults to help further the theoretical understanding of the spacing effect in group settings across development.

The above discussion shows that applying laboratory-developed theories to classroom training can be inappropriate without further testing. Thus, as spaced retrieval practice has received extensive laboratory testing but has not been matched by classroom-based testing, we need to conduct the experiments suggested in the above sections before recommending it for use in classroom settings.

## 3.6 Conclusion

The current experiments examined whether within-day spaced retrieval practice effectively promotes retention of novel animal names one week after initial exposure. Our results were surprisingly in contrast with extant literature in that we did not observe a clear spacing benefit at any test point. Without an additional retrieval opportunity on Day 2, we observed that within-day spaced retrieval practice led to further improvements in word learning performance one week later, which was not observed after massed retrieval practice. However, this was not enough to elicit a clear spacing benefit on Day 7. The current theories cannot explain our findings, highlighting the need for continued research of spaced retrieval practice, especially in schools, before promoting it as a viable method in educational settings. Further, there is a lack of studies directly comparing the spacing effect in adults and children, meaning that we cannot determine whether the same theories underpin the spacing effect or if it is, in fact, the same effect in adults and children.

# Chapter 4:

## A direct comparison of the effects of within-day spaced retrieval practice in adults and children

The data and analysis outputs for all experimental comparisons are available at the Open Science Framework: https://osf.io/tvcm9/

The pre-registration for Analysis 2 (Experiments 3 and 5 comparisons) is available at the Open Science Framework: https://osf.io/s4azr

## 4.1 Abstract

When examining the spaced retrieval practice literature, there are tentative developmental differences. Specifically, studies have found an earlier emerging spacing benefit in children, while adults show a benefit from massed retrieval practice. However, there is a lack of studies directly comparing the spacing effect in adult and child populations. To close this gap in the literature, we conducted two cross-experiment analyses to determine whether within-day spaced, and massed retrieval practice affects word learning differently for adults and children. The results indicated that the initial retrieval practice schedule only affected retention one week later differently depending on the age of the participants if a retrieval opportunity was given after a period of sleep. Namely, adults who practised retrieval in a massed schedule maintained their performance, while adults in the spaced schedules and children in all schedules continued to improve from the first to the second day, but only children continued to improve one week later. This finding suggests that sleep-based consolidation and reconsolidation must be considered in spaced retrieval practice, especially when considering its application in different populations.

## 4.2 Introduction

In the previous chapters, we have reported somewhat surprising findings; the lack of a spacing benefit for either adults or children was not expected based on previous literature. However, we observed an effect of the spaced and massed schedules in adults and children, meaning that it is important to determine whether the current populations showed the same effect pattern from within-day spaced retrieval practice or not. Thus, this chapter will first outline how the word learning can differ between adults and children, especially regarding sleep-based consolidation, and then whether extant literature report differences in the effect of spaced retrieval practice. We will then analyse the data from Chapters 2 and 3 to compare how within-day spaced retrieval practice directly may interact with word learning in adults and children and if the effects shown were different depending on the age of the participants. As mentioned in Chapter 3, there is a severe lack of studies directly comparing the effect of spaced retrieval practice in adults and children. Thus, this chapter will provide a strong opening for furthering our understanding of how word learning through spaced retrieval practice can act in different populations and whether theoretical accounts can be applied to the spacing benefit regardless of age.

As outlined in Chapter 1, prior knowledge and sleep-based consolidation processes can affect word learning differently in adults and children. Thus, it is important to closely examine these aspects in the context of spaced retrieval practice to determine if the method can be applied across all developmental stages with the same outcome. The studies below all report a spacing benefit (i.e., better recall of novel words/word pairs) in long-term measures, taking place between 1 and 14 days after initial learning. However, as we will show, they all report different performance patterns during the initial retrieval practice sessions, depending on the age of the participants.

First, as we reported in Chapter 2, adults showed an initial benefit in word learning from massed retrieval practice, as steeper improvements across test sessions on the first day. This is a relatively common pattern in the existing literature (e.g. Miles, 2014; Cepeda et al., 2009; Karpicke & Roediger, 2007) and could be argued to show adults' ability to use prior knowledge (i.e. already known words) to aid rapid initial learning. This line of argument could explain why an initial massed benefit can emerge in adult populations regardless of whether a sleep period occurs between retrieval practice sessions. For example, Karpicke and Roediger (2007) reported that adults learning novel words in massed presentations could initially recall more words than if the words were learned in spaced presentations (i.e. no sleep occurred between retrieval attempts). Similarly, Cepeda et al. (2009) found an initial massed benefit when retrieval practice was spaced over one or multiple days (up to 14 days apart). This could indicate that adults rely on prior knowledge to support initial learning more than on sleep-based consolidation between spaced sessions.

In contrast, children often show an initial spacing benefit that maintains in long-term measures, but mainly if a sleep period occurs between retrieval occasions. For example, children learning novel word pairs in immediate presentations (massed) or spaced 5 or 10 presentations apart could recall an equal amount of word pairs at the end of an initial encoding phase (Zigertman, Simone & Bell, 2015). In Chapter 3, we also showed that massed retrieval practice resulted in higher performance than spaced retrieval practice before sleep. In contrast, Moinzadeh et al. (2008) taught children aged 12-13 unfamiliar English words in a longer massed session on a single day or in shorter sessions spaced one day apart, and immediately after the training sessions, children in the spaced condition could recall significantly more words than those in the massed condition. Similarly, Goossens et al. (2012) reported better initial retention of novel words and their definitions if they were spaced rather than massed. This could suggest that children, who have less prior knowledge to bind new information to, rely on their superior sleep-based consolidation ability to aid learning, resulting in an initial spacing benefit.

In short, there are both similarities (i.e., long-term spacing benefit reported in both adult and child populations) and differences (i.e., an initial benefit from massed practice in adults versus an initial spacing benefit in children) in the spacing literature. However, while it is tempting to conclude that developmental differences are present in the spacing effect, especially in the initial learning phase, we must be cautious in making claims like this. The studies above are individual studies with different methodologies, making direct comparisons potentially inaccurate. For example, the adult studies (Miles et al., 2014; Cepeda et al., 2009) are based on individual testing, while some child studies are based on classroom testing (e.g. Goossens et al., 2012; Moinzadeh et al., 2008). The studies also use different types of tasks and stimuli, which could tap into different learning mechanisms, resulting in different learning outcomes. These methodological deviations highlight the importance of direct comparisons of adult and child populations using the same paradigms and methods, an area currently lacking in the literature.

### 4.2.1 The current analyses

To address the lack of studies in the current literature directly comparing the spacing effect in adult and child populations, we conducted two cross-experiment analyses to determine whether within-day spaced, and massed retrieval practice interacts with word learning differently for adults and children. First, we compared the adult and child data from

Experiments 2 and 4 to determine how within-day spaced retrieval practice, and a follow-up session after sleep affects word learning on Day 1 and retention on Day 2 and one week later (Day 7).

We cannot base our expectations for these analyses solely on extant literature due to the lack of direct comparisons between adult and child studies. Thus, we based our expectations for the following analyses on our results in Experiments 2-5.

The first set of analyses will explore our findings of Experiments 2 and 4, providing a developmental comparison of within-day spaced retrieval practice effects on word learning performance before sleep, 24 hours, and one week later. On Day 1, based on the Schedule*Block effect observed in our adult data (Experiment 2) and the trend towards the same in the child data (Experiment 4), we expected a similar effect in the current analyses. Specifically, we hypothesised that we would see steeper improvements in our word learning tests (cued recall, picture naming) across blocks on Day 1 (i.e., a Schedule*Block interaction). Additionally, as adults received an additional retrieval practice opportunity (see Table 8 for an outline of the key methodological differences between the experiments), we expected them to improve their word learning performance at a steeper rate than children across blocks on Day 1 (seen as an Age*Block interaction). Finally, in the overnight and long-term performance analyses, we expected children to improve their performance at a steeper rate than adults across the days (seen as an Age*Day interaction; based on existing literature suggesting enhanced benefits from overnight consolidation processes in children, e.g., James et al., 2019; Wilhelm et al., 2012). We also expected adults and children to change their performance across days at different rates, depending on their Schedules. This was based on the finding of a

Schedule*Block interaction in Experiment 2 (i.e., adult data), but a lack of interaction in Experiment 4 (i.e., child data).

The second set of analyses explored the findings of Experiments 3 and 5, directly comparing the effects of within-day spaced retrieval practice on the retention of word retention one week later in adults and children (analyses and hypotheses were preregistered as part of Experiment 5; https://osf.io/s4azr). Based on our previous findings, we preregistered our expectations as:

1) Across blocks on Day 1, both Adults and Children would show steeper improvements if they practised retrieval in a Massed session compared to Spaced sessions (i.e., support for a Schedule*Block interaction, but no support for an Age*Schedule*Block interaction).
2) Children would show a greater effect from Spaced retrieval practice (compared to Massed retrieval practice) in overall changes in long-term measures (i.e., from Day 1 to Day 7) relative to Adults, seen as support for an Age*Schedule*Block interaction in the analysis.

## 4.3 Methods

### 4.3.1 Data preparation

While we aimed to keep the experiments as similar as possible, there were minor differences between the methods. The key methodological differences are outlined in Table 8, and before running the analyses, we would like to focus on the difference in the number of retrieval blocks on Day 1. Due to scheduling constraints in Experiments 4 and 5, we only have data from two timepoints on Day 1 in the Child data, compared to three points in the Adult data. Therefore, before analysing the data on Day 1, we had to decide which time points to use. As we were mainly interested in overall performance changes across Day 1, we decided to compare performance from the first Block (i.e. B1 in both the adult and child data) and the final Block on Day 1 (i.e. B3 in the adults, B2 in the children).

**Table 8.** Methodological differences in the experiments used in the adult-child comparisons analyses.

| | **ANALYSIS 1** | | **ANALYSIS 2** | |
|---|---|---|---|---|
| | **Experiment 2 (Adults)** | **Experiment 4 (Children)** | **Experiment 3 (Adults)** | **Experiment 5 (Children)** |
| **No. of retrieval blocks on Day 1** | 3 | 2 | 3 | 2 |
| **Total no. of retrieval blocks** | 5 | 4 | 4 | 3 |
| **No. of stimuli** | 20 | 12 | 20 | 12 |
| **Experimental setting** | Laboratory (individual) | Classroom (group) | Online (individual) | Classroom (group) |

### 4.3.2 Statistical analyses

Like our previous analyses, we conducted our analyses in JASP version 0.16 (JASP Team, 2021). We also used the same criterion for BF support, where a BF of 1 indicates no support for either hypothesis, BF between 1 and 3 are weak support, BF between 3 and 10 are considered moderate, and values above ten are strong support (van Dorn et al., 2020; Kass & Raftery, 1995; Jeffreys, 1939). Values below 0.3 indicate support for the null hypothesis.

We used the matching percentage (based on the Levenshtein Distance) as the analysis's dependent variable. The below sections outline which analyses we conducted for the two comparisons. We only compared data from the Cued Recall and Picture Naming tests, as the

Base Animal Match data in the Child studies were unsuitable for analysis (see Chapter 3, Section 3.3.2, for a reason for this decision).

#### *4.3.2.1 Analysis 1: comparing the effect of within-day spaced retrieval practice on retention of novel words before and after sleep, and one week later in adults and children*

To compare performance changes across Day 1 in Experiments 2 and 4, we conducted a 2(Age: Adult, Child) x 3(Schedule: Massed AM, Massed PM, Spaced) x 2(Block: First block on Day 1, Final block on Day 1) Bayesian repeated-measures ANOVA. Then, to compare performance changes from Day 1 to Day 2 and Day 7, we conducted a 2(Age: Adult, Child) x 3(Schedule: Massed AM, Massed PM, Spaced) x 3(Day: Final Block on Day 1, Day 2, Day 7) Bayesian repeated-measures ANOVA.

#### *4.3.2.2 Analysis 2: comparing the effect of within-day spaced retrieval practice on novel word retention one week after initial training*

To compare performance changes across Day 1 in Experiments 3 and 5, we conducted a 2(Age: Adult, Child) x 2(Schedule: Massed, Spaced) x 2(Block: First block on Day 1, Final block on Day 1) Bayesian repeated-measures ANOVA. Then, we conducted a 2(Age: Adult, Child) x 2(Schedule: Massed, Spaced) x 2(Day: Final Block on Day 1, Day 7) Bayesian repeated-measures ANOVA to determine whether performance changes from Day 1 to Day 7 was affected by Age and/or Schedule.

## 4.4 Analyses results

### 4.4.1 Analysis 1 Results

In Analysis 1 we compared the data from Experiments 2 and 4 to determine whether performance in word learning tests in adults and children was affected by within-day spaced or massed retrieval practice on the day of practice (Day 1), one day (Day 2) and one week (Day 7) later.

#### *4.4.1.1 Changes on Day 1*

In the first analysis, there was strong support for a main effect of Block (cued recall $BF_{10}$=1.1e+23, error%=0.88; picture naming $BF_{10}$=7.9e+28, error%=1.16) and Age in the picture naming task (picture naming $BF_{10}$=5.90, error%=1.63) but not in the cued recall test ($BF_{10}$=0.22, error%=1.22) or for a main effect of Schedule (cued recall $BF_{10}$=0.42, error%=0.70; picture naming $BF_{10}$=0.71, error%=0.94). Overall, adults performed higher than children in the picture naming task, but not in the cued recall test, and overall performance improved across Blocks on Day 1. However, there was no overall difference in performance between participants in the different schedules.

Further, there was overwhelming support for a Block*Schedule interaction (cued recall $BF_{incl}$=38343.24; picture naming $BF_{incl}$=4128.3) and Block*Age (cued recall $BF_{incl}$=7.7e+10; picture naming $BF_{incl}$=3.9e+15). As seen in Figure 12, overall, participants in the two Massed schedules improved their recall levels of the animal names at a steeper rate than participants in the Massed schedule. Further, the support for the Block*Age interaction (illustrated in Figure 13) suggests that adults improved their performance at a steeper rate than children. To determine whether the steeper rate of improvement across blocks for adults was due to the age variable or simply because adults completed one more round of retrieval practice, we conducted a further analysis comparing performance changes from the first (B1) to the second (B2) block for both adults and children. The analysis maintained strong support for a Block*Age interaction (cued recall $BF_{incl}$=83.78, picture naming $BF_{incl}$=263.63), indicating that the differences in performance change across Blocks on Day 1 were related to the Age of participants.

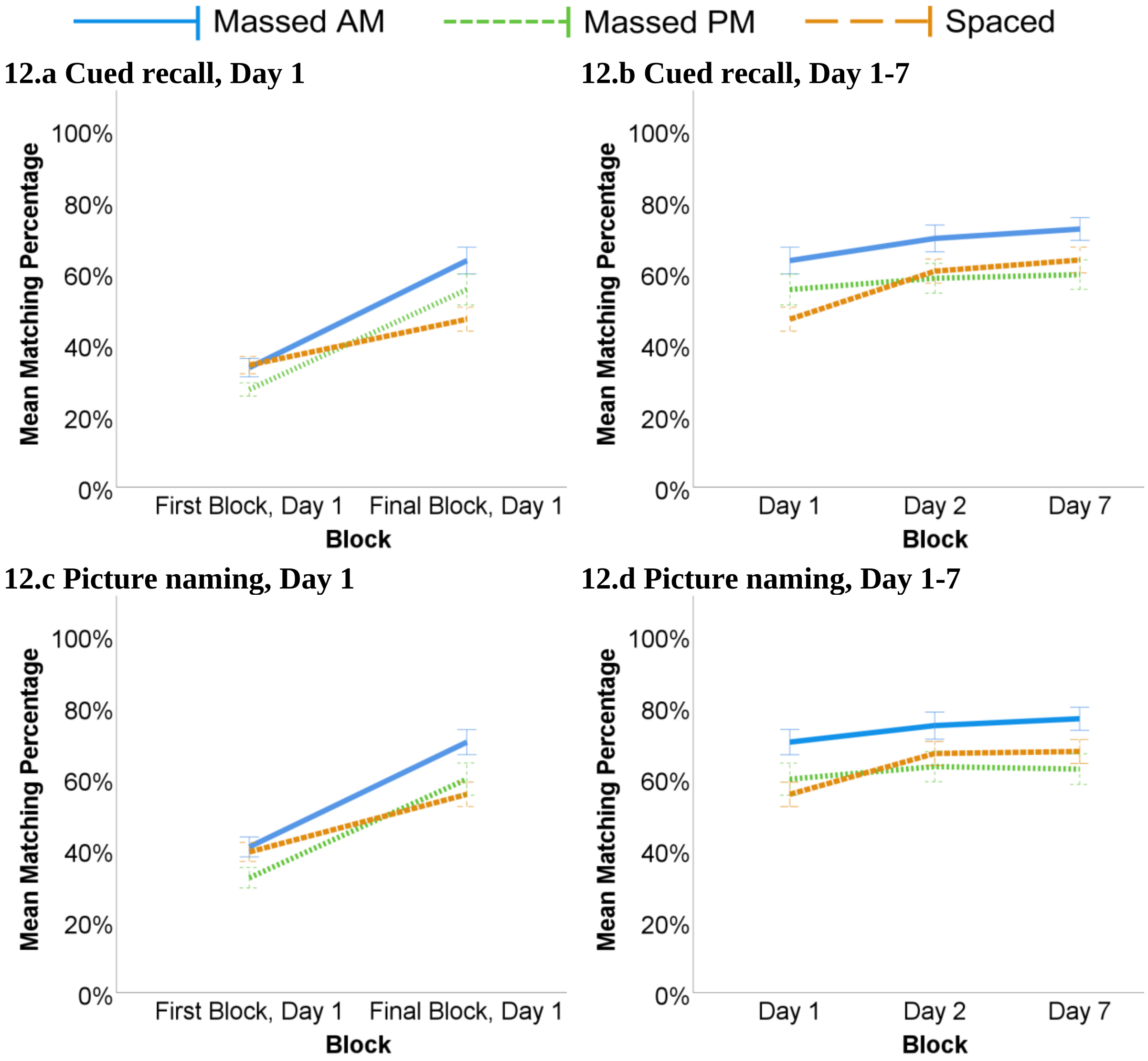
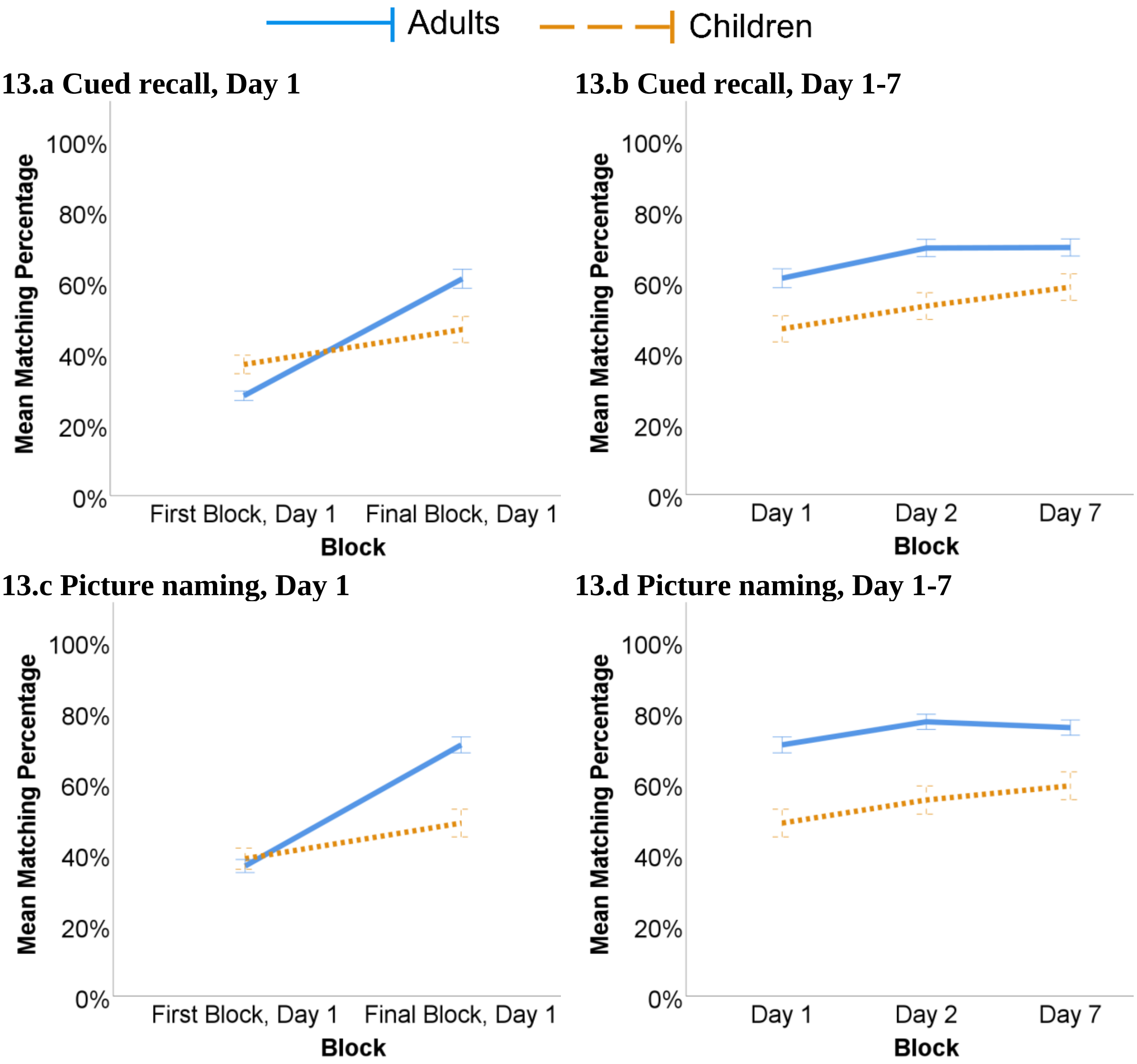


**Figure 12. Results of Adult-Child comparison analyses (Experiments 2 and 4), focused on the Schedule interactions.** Illustrating Schedule*Block and Schedule*Day interactions for Cued Recall Day 1 (**12.a**) and Day 1-7 (**12.b**) and Picture Naming Day 1 (**12.c**) and Day 1-7 (**12.d**) tests in the adult and child analyses.

There was no support for the Age*Schedule interaction (cued recall $BF_{incl}$=2.96; picture naming $BF_{incl}$=1.07) and no support for the Age*Schedule*Block interaction (cued recall $BF_{incl}$=1.25, picture naming $BF_{incl}$=0.45). The lack of support for these interactions shows no evidence that adults and children performed differently depending on their schedules. In short, this analysis shows that adults can learn novel words through retrieval practice with feedback faster than children, indicating developmental differences before a period of sleep. Further, adults and children learned the novel animal names faster if they practiced retrieval with feedback in massed sessions than if a temporal space separated them, meaning there was no evidence for developmental differences in the spacing effect before sleep.



**13.a Cued recall, Day 1**

**13.b Cued recall, Day 1-7**

**13.c Picture naming, Day 1**

**13.d Picture naming, Day 1-7**

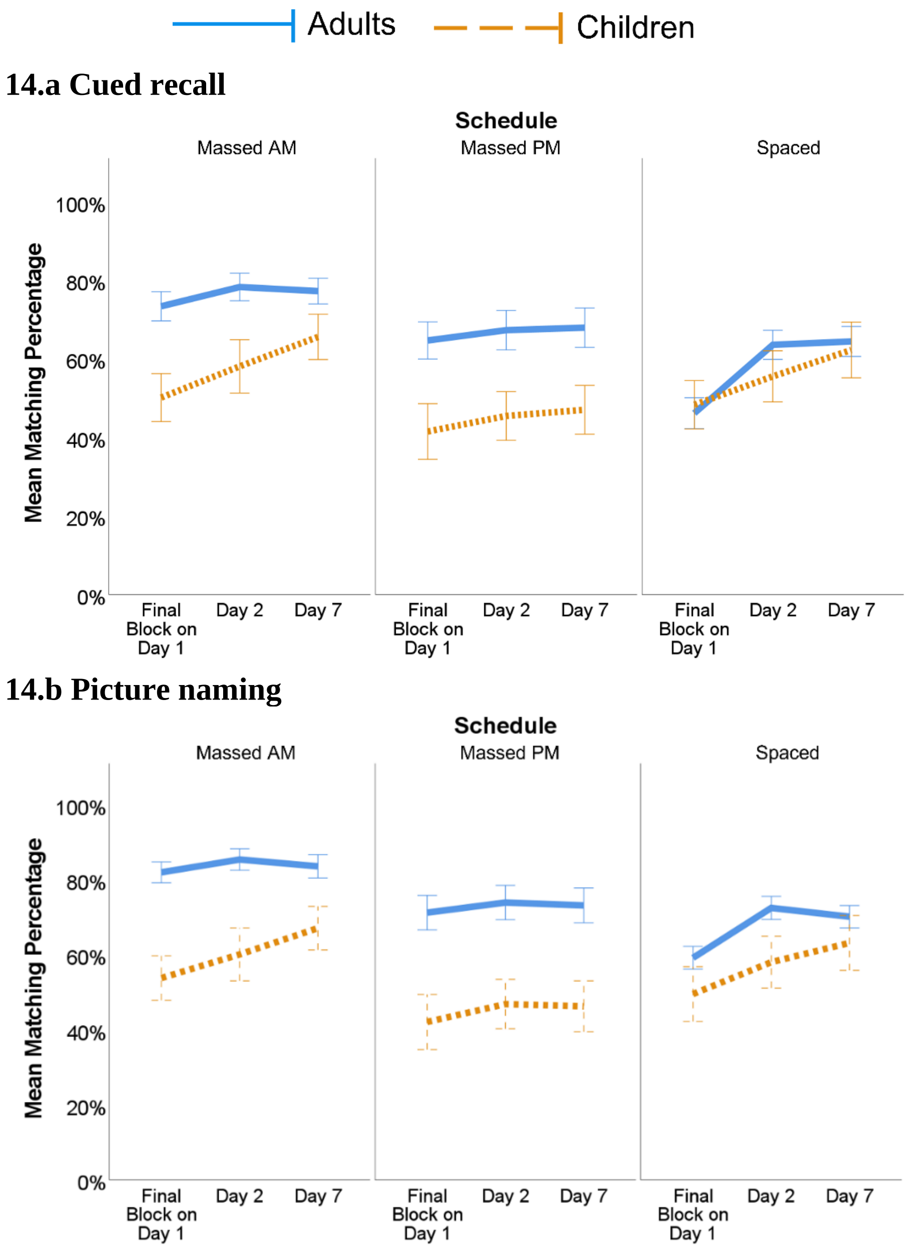
**Figure 13. Results of Adult-Child comparison analyses (Experiments 2 and 4), focused on the Age interactions.** Illustrating Age*Block and Age*Day interactions for Cued Recall Day 1 (**13.a**) and Day 1-7 (**13.b**) and Picture Naming Day 1 (**13.c**) and Day 1-7 (**13.d**) tests in the adult and child analyses.

#### *4.4.1.2 Changes from Day 1 to Day 2 and Day 7*

Focusing on changes from Day 1 to Day 2 and Day 7, there was strong support for a main effect of Age (cued recall $BF_{10}$=26.74, error%=1.87; picture naming $BF_{10}$=4830.01, error%=4.08), and for Day (cued recall $BF_{10}$=7.3e+15, error%=0.66; picture naming $BF_{10}$=7.0e+14, error%=0.62), but there was no support for a main effect of Schedule (cued recall $BF_{10}$=1.65, error%=7.04; picture naming $BF_{10}$=1.38, error%=4.11). These results indicate that adults performed overall higher than children and that the overall performance improved from Day 1 to Day 2 and Day 7. However, the overall performance in the different schedules did not differ.

The analysis revealed overwhelming support for a Day*Schedule interaction in both tests (cued recall $BF_{incl}$=1609.12; picture naming $BF_{incl}$=1087.2), but there was no support for an Age*Schedule interaction in cued recall ($BF_{incl}$=0.94) and no support in the picture naming test ($BF_{incl}$=0.55). As seen in Figure 12, participants in the Massed schedules overall maintained performance while participants in the Spaced schedule improved performance from Day 1 to Day 2 and then maintained their performance to Day 7.

Interestingly, there was no support for an Age*Day interaction in the cued recall test ($BF_{incl}$=0.82) but strong support for the interaction in the picture naming test ($BF_{incl}$=67.12). Figure 13 shows that adults' performance in the picture naming test was a trend toward maintenance compared to children who showed consistent improvements. In contrast, in the cued recall test, adults improved performance from Day 1 to Day 2, which was more similar to the children's performance.



14.a Cued recall

14.b Picture naming

**Figure 14. Results of Adult-Child comparisons (Experiments 2 and 4), Day 1-7 analysis.** Figures show the mean matching percentage across blocks on Day 1 to Day 2 and Day 7 in the Cued Recall (**14.a**) and Picture Naming (**14.b**) tests. Blue solid bars represent Adult data; Orange dashed bars represent Child data. Error bars represent +-1SE.

Finally, there was strong support for an Age*Schedule*Day interaction in the cued recall test ($BF_{incl}$ = 7.48) but no support for the interaction in the picture naming test ($BF_{incl}$ = 2.67). The interaction in the cued recall task is important as this brings evidence that overnight and long-term retention of novel words one week later was affected by within-day spaced or massed retrieval practice differently depending on the age of participants (see Figure 14). However, the interaction failed to reach statistical significance in the picture naming task, suggesting that performance in the cued recall task (i.e., orthographic memory) was more sensitive to differences in Age and Schedule across the Days than in the picture naming task (i.e., associative memory). This will be further discussed in 4.5 Discussion.

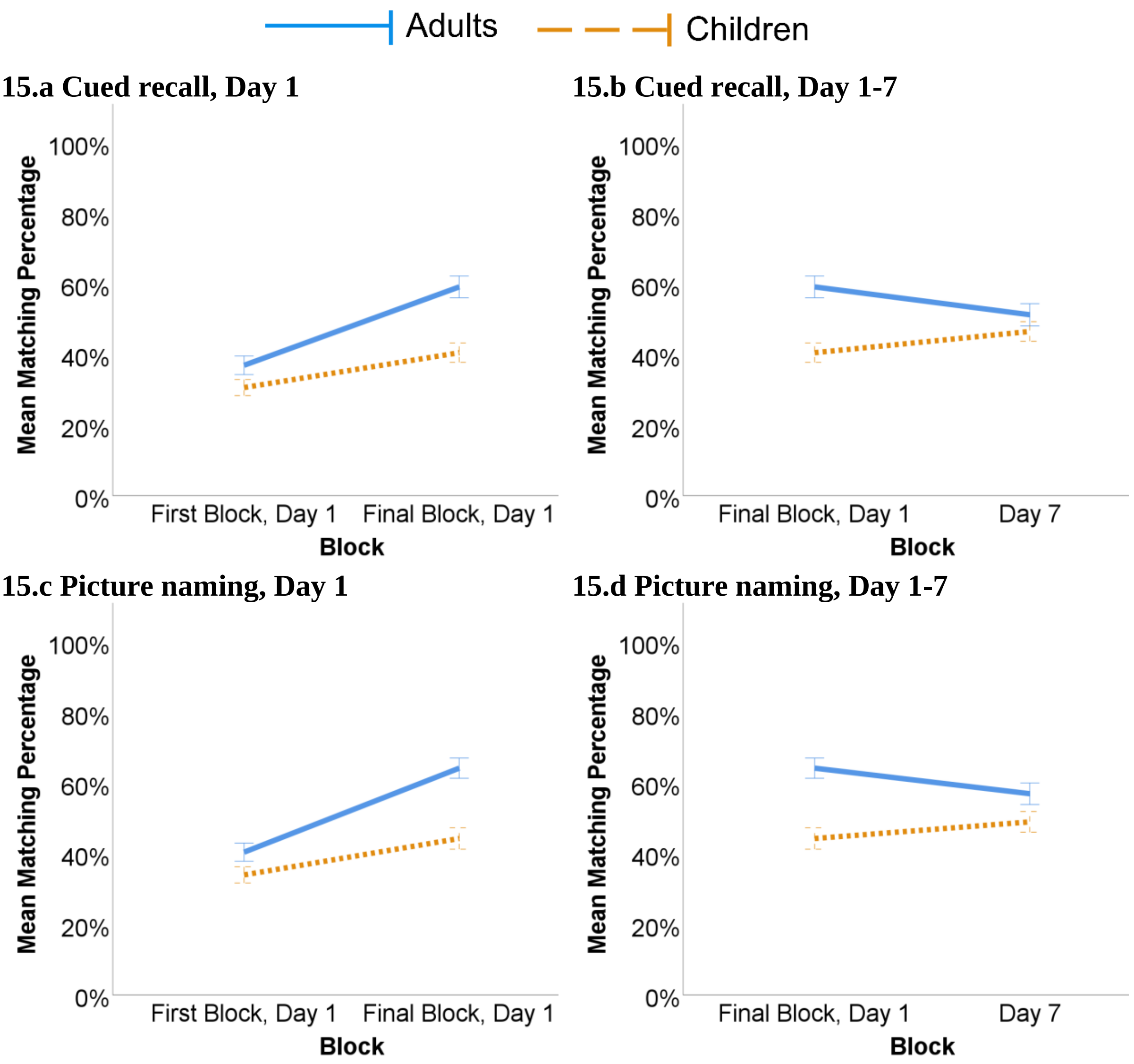
### 4.4.2 Analysis 2 Results

In Analysis 2, we focused on potential developmental differences in retention of novel words one week after initial exposure and whether initial retrieval practice schedules affected this.

#### *4.4.2.1 Changes on Day 1*

In the first analysis, there was strong support for a main effect of Block (cued recall $BF_{10}$=6.5e+14, error%=1.11; picture naming $BF_{10}$=2.1e+14, error%=1.42) and Age (cued recall $BF_{10}$=44.30, error%=2.08; picture naming $BF_{10}$=96.4, error%=0.93) but not for Schedule (cued recall $BF_{10}$=0.86, error%=1.11; picture naming $BF_{10}$=0.53, error%=0.83). Overall, adults performed higher than children, and performance improved across blocks on day 1. However, there was no overall difference in performance between participants in the different schedules.

Further, there was strong support for the Age*Block (cued recall $BF_{incl}$=883.1; picture naming $BF_{incl}$=414.2) and Schedule*Block (cued recall $BF_{incl}$=8189.2; picture naming $BF_{incl}$=18.83) interactions. Like in Analysis 1, adults improved at a steeper rate than children across blocks on Day 1 (as illustrated in Figure 15), and participants in the Massed schedule improved at a steeper rate than those in the Spaced schedule (illustrated in Figure 16).



15.a Cued recall, Day 1

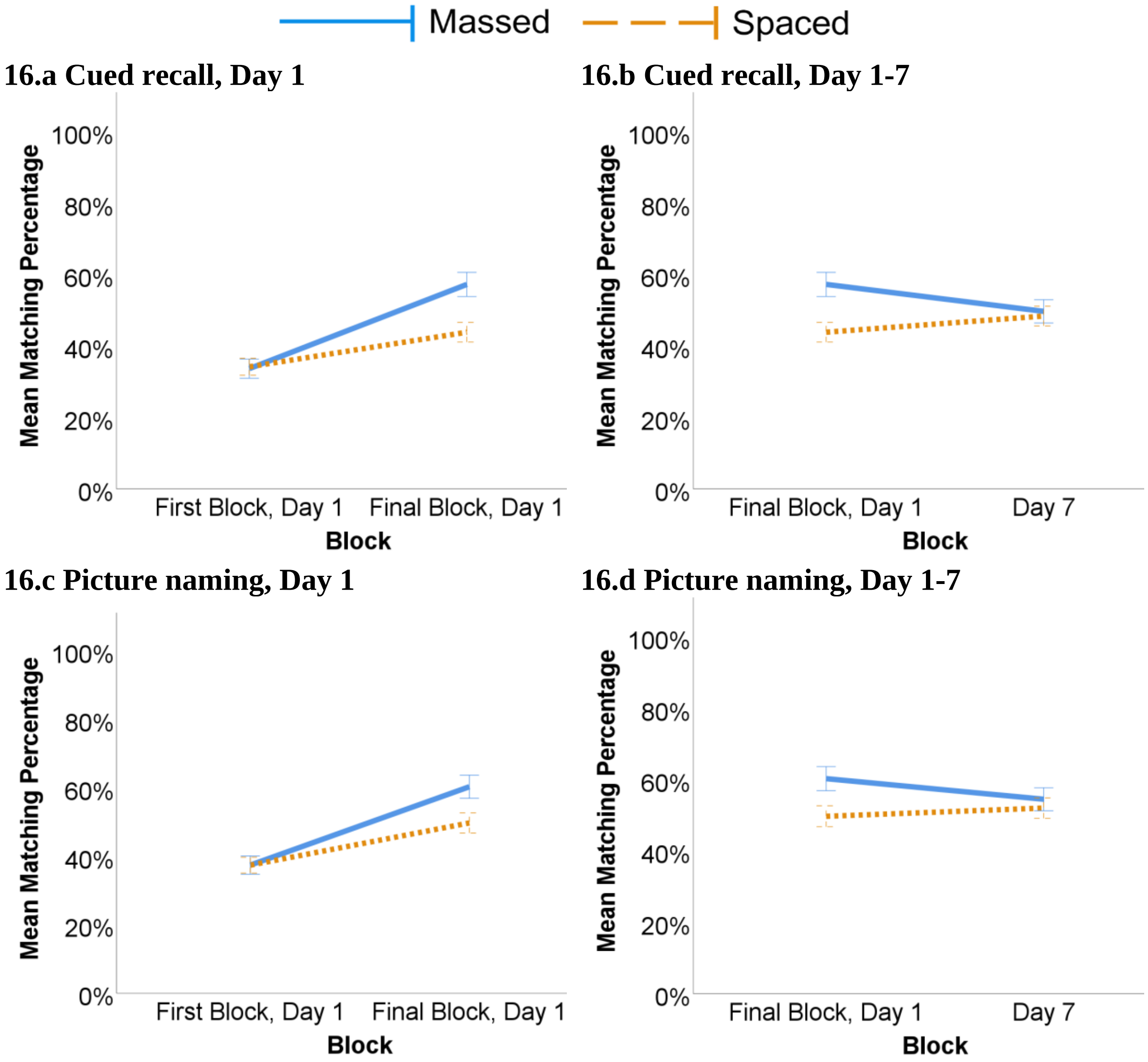
15.b Cued recall, Day 1-7

15.c Picture naming, Day 1

15.d Picture naming, Day 1-7

**Figure 15. Results of Adult-Child comparison analyses (Experiments 3 and 5), focused on Age interactions.** Illustrating Age*Block and Age*Day interactions for Cued Recall Day 1 (**15.a**) and Day 1-7 (**15.b**) and Picture Naming Day 1 (**15.c**) and Day 1-7 (**15.d**) tests in the adult and child analyses.

There was no support for the Age*Schedule interaction (cued recall BF$_{incl}$=0.45; picture naming BF$_{incl}$=0.36) and weak support for the null hypothesis in the Age*Schedule*Block interaction in both tests (cued recall BF$_{incl}$=0.37; picture naming BF$_{incl}$=0.31). This was in line with our expectations and indicated that both Adults and Children improved at a steeper rate in Massed Schedules compared to Spaced.



**16.a Cued recall, Day 1**

**16.b Cued recall, Day 1-7**

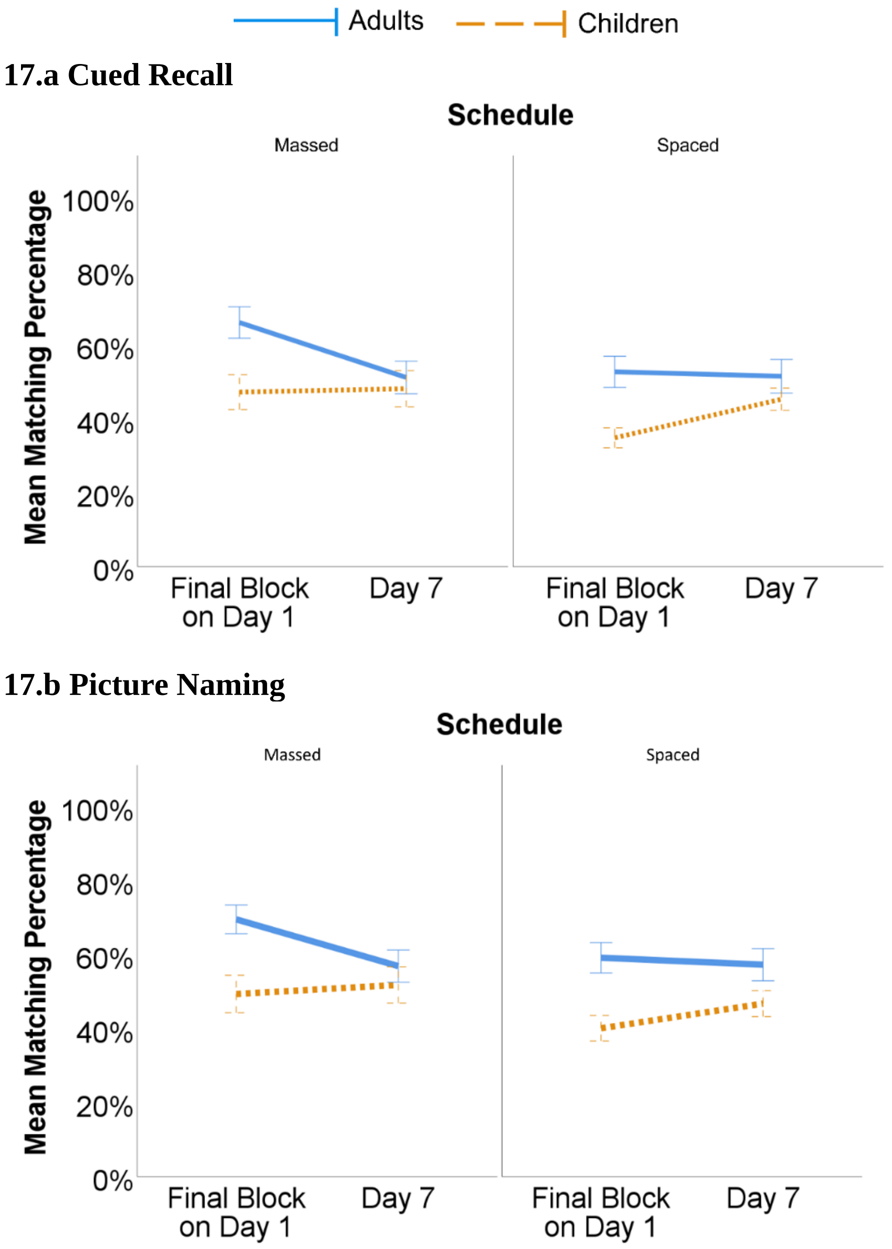
**16.c Picture naming, Day 1**

**16.d Picture naming, Day 1-7**

**Figure 16. Results of Adult-Child comparison analyses (Experiments 3 and 5), focused on Schedule interactions.** Illustrating Schedule*Block and Schedule*Day interactions for Cued Recall Day 1 (**16.a**) and Day 1-7 (**16.b**) and Picture Naming Day 1 (**16.c**) and Day 1-7 (**16.d**) tests in the adult and child analyses.

#### *4.4.2.2 Changes from Day 1 to Day 7*

Focusing on changes from Day 1 to Day 7, there was again strong support for a main effect of Age (cued recall $BF_{10}$=10.10, error%=1.38; picture naming $BF_{10}$=33.92, error%=1.67), but no support for a main effect of Schedule (cued recall $BF_{10}$=1.13, error%=1.56; picture naming $BF_{10}$=0.89, error%=2.75) and support for the null hypothesis for a main effect of Day (cued recall $BF_{10}$=0.20, error%=5.06; picture naming $BF_{10}$=0.28, error%=1.25). These results again indicate that adults performed overall higher than children and that the overall performance in the different schedules did not differ. Additionally, there was weak support for the null hypothesis in the Day variable, indicating that the overall performance levels were not different on Day 1 and Day 7.



17.a Cued Recall

17.b Picture Naming

**Figure 17. Results of Adult-Child comparisons (Experiments 3 and 5), Day 1-7 analysis.** Figures show the mean matching percentage across blocks on Day 1 to Day 2 and Day 7 in the Cued Recall (**17.a**) and Picture Naming (**17.b**) tests. Blue solid bars represent Adult data; Orange dashed bars represent Child data. Error bars represent +-1SE.

While the analysis revealed overwhelming support for the Age*Day interaction (cued recall $BF_{incl}$=17597.0; picture naming $BF_{incl}$=7421.0) and strong support for Schedule*Day (cued recall $BF_{incl}$=1143.2; picture naming $BF_{incl}$=22.08), there was no support for an Age*Schedule (cued recall $BF_{incl}$=0.57; picture naming $BF_{incl}$=0.52) or Age*Schedule*Day interaction (cued recall $BF_{incl}$=0.34; picture naming $BF_{incl}$=0.66). Figure 15 shows that changes in performance from the final Block on Day 1 depended on Age (Children showed overall improvements while Adults showed overall forgetting) and Schedule (Spaced showed overall improvements and Massed overall forgetting, illustrated in Figure 16), but not the two variables combined (see Figure 17).

## 4.5 Discussion

In this chapter, we have directly compared how adults and children benefit from spaced and massed retrieval practice in word learning tests. Our analyses revealed that, on the first day, adults improved their word retention performance at a steeper rate than children. We further showed that this was due to the age difference, not due to adults receiving an additional retrieval opportunity. We also showed that both adults and children benefited from massed retrieval practice more than spaced, as they learned more animal names before sleep. These patterns were present in both analyses, supporting that adults and children do not differ in the effects of within-day spaced and massed retrieval practice before sleep.

Interestingly, after a sleep period, our analysis revealed support for an interaction between how adults and children change their performance depending on which group they were in. However, statistical support for this only emerged if an additional retrieval opportunity was given the day after initial practice (see Analysis 1). Specifically, adults who practised retrieval in within-day spaced sessions improved their recall performance from the first to the second day and then maintained this one week later. Instead, adults who practised massed retrieval maintained their already higher performance, while children improved their performance continually from the first to the second day and one week later. This indicated that developmental differences could be present in the spacing effect. In contrast, and counter to our expectations for analysis 2, we did not observe developmental differences depending on the initial retrieval schedule if no retrieval opportunity was provided the day after initial practice. However, we found that adults showed overall forgetting of the animal names one week after initial practice, while children showed improvements in their performance. We also observed that participants practising retrieval in an initial within-day spaced schedule maintained their performance, while participants practising retrieval in a massed schedule showed overall forgetting one week later. Together, these results indicate that within-day spaced retrieval practice protected memories from forgetting the same for both adults and children, but developmental differences emerged once an additional retrieval opportunity was provided after one night's sleep.

An interesting and important finding of the current analyses is that we found strong support that the initial retrieval practice schedule and age affected retention levels one week after initial exposure, but only when a follow-up session took place after a sleep period. Specifically, if no retrieval opportunity occurred one day after initial exposure, adults showed overall forgetting, and children improved, but this was not affected by the initial retrieval practice schedule. However, once a retrieval opportunity was provided 24 hours after exposure, the trajectory of change in retention of the novel animal names differed depending on the initial retrieval practice schedule for adults and children. As argued by Smith and Scarf (2017), retrieval that occurs after a period of consolidation (e.g., during the period of overnight sleep between the first and second day of Experiments 2 and 4) could place the consolidated memory traces in a fragile state, allowing for alteration and further strengthening of the memory traces in subsequent offline consolidation periods. We observed a developmental difference in the spacing effects once retrieval practice followed a period of offline consolidation, but not if no reconsolidation opportunity took place, providing support for the reconsolidation account by Smith and Scarf (2017).

Interesting, we only observed strong support for the above developmental difference in the cued recall test but not in picture naming. Cued recall mainly measures orthographic memory, while picture naming measures associative memory. Thus, the finding of a developmental difference in performance changes depending on the initial retrieval practice schedule in the cued recall test indicates that the memory types may be differently affected by the spacing effect and a reconsolidation opportunity. Many studies exploring the effects of spaced retrieval practice on word learning use paired associates or word-pair paradigms (e.g., Kroneisen & Kuepper-Tetzel, 2021; Bell et al., 2014; Sobel, Cepeda & Kapler, 2011; Karpicke & Roediger, 2010; Pyc & Rawson, 2009; Logan & Balota, 2008; Karpicke & Roediger, 2007), which are similar to our picture naming test. However, few studies explore orthographic memory (e.g., Lindsay & Gaskell, 2013; Moss, 1996; Rea & Modigliani, 1985), and even fewer consider spelling changes (Bloom & Shuell, 1981) when assessing retention levels. Our results demonstrate that orthographic memory performance after spaced or massed retrieval practice is susceptible to developmental differences if a reconsolidation opportunity is given, highlighting the importance of examining this further.

## 4.6 Conclusion

To summarise, the current cross-experiment comparisons aimed to determine whether adults and children showed different effects from within-day spaced retrieval practice with and without a reconsolidation opportunity one day later. The results indicate that the initial retrieval practice schedule only affected retention one week later differently depending on the age of the participants if a retrieval opportunity was given after a period of sleep. Children continued to improve their word learning performance from Day 1 to Day 2 and Day 7, regardless of the initial retrieval schedule if a post-sleep retrieval opportunity was provided. Adults only showed continued improvements from Day 1 to Day 2 if their initial retrieval practice was spaced but maintained performance from Day 2 to 7, while adults in the massed schedules maintained performance from Day 1 to Day 7. Thus, a post-sleep retrieval opportunity impacted later retention of novel words differently depending on the participant's age and initial retrieval schedules. This finding enhances the argument made by Smith and Scarf (2017) that the role of sleep-based consolidation and reconsolidation needs to be considered in the context of spaced retrieval practice. Further, this highlights the need for future studies to carefully explore developmental differences in the effects of spaced retrieval practice and how it might influence theoretical accounts.

# Chapter 5:

## General Discussion

This thesis has examined two methods for encouraging long-term retention of novel words and whether developmental differences in underlying word learning mechanisms affect their effectiveness. Specifically, we have explored whether within-day spaced or massed retrieval practices effectively promote learning and memory of unfamiliar animal names in adults and children. Spacing retrieval practice over separate days has been consistently reported to result in better word learning performance in long-term measures compared to massed retrieval practice (i.e. when all practice takes place in a single session on a single day). However, the spacing effect in the literature shows similarities to sleep-based consolidation and repeated retrieval practice with feedback, yet relatively few studies have examined these effects.

This thesis addresses three critical empirical issues to advance this literature. First, according to the CLS account of word learning, sleep can greatly impact the retention of novel words through sleep-based consolidation, and despite the wealth of evidence supporting this position across development, the role of sleep in spaced retrieval practice has not been thoroughly considered. In fact, there are several studies where a reported spacing effect could potentially be attributed to sleep rather than spacing *per se*. Here, we addressed this by spacing retrieval practice sessions at least 2 hours apart within a single day, therefore removing the potential influence of intervening sleep. Second, feedback is often included in studies exploring spaced retrieval practice, muddying the distinction between spaced *learning* and spaced *retrieval practice*. Based on theoretical accounts of spaced retrieval practice, we examined the critical assumption that a spacing effect will emerge regardless of whether intervening sleep or feedback was included, regardless of the age of participants. Third, despite the assumption that both children and adults show a spaced retrieval practice benefit, with this feeding into educational practice, there have been no direct developmental comparisons despite some evidence of subtle developmental differences in the time course of word learning. Here we carry out the first direct child-adult comparison of spaced retrieval practice in the context of word learning without sleep between retrieval attempts.

To address these key issues, through five experiments, we tested whether a spacing benefit emerged after within-day spaced retrieval practice without further learning opportunities through feedback (Experiment 1, Chapter 2) and with feedback allowing additional exposure during retrieval attempts in adults (Experiment 2, Chapter 2) and children (Experiment 4, Chapter 3). Furthermore, we examined whether a spaced retrieval practice benefit would emerge without a post-sleep retrieval opportunity in both massed and spaced conditions in adults (Experiment 3, Chapter 2) and children (Experiment 5, Chapter 3). Finally, to determine whether adults and children showed the same effect from within-day spaced retrieval practice, potentially allowing theoretical accounts to be advanced across development, we directly compared the child and adult data from the above experiments (Chapter 4). We showed that the spaced retrieval practice effect is not as robust as theoretical accounts may infer and that critical factors such as intervening sleep and feedback need to be considered in future theoretical accounts.

This General Discussion chapter will first summarise the key findings of the current experiments before discussing the broader theoretical implications of feedback and sleep/reconsolidation in the spacing effect from a developmental perspective.

## 5.1 Summary of experimental findings

### 5.1.1 Chapter 2: The effect of feedback and post-sleep retrieval on word learning after within-day spaced and massed retrieval practice in adults

We investigated whether within-day spaced retrieval practice with and without feedback leads to a benefit in immediate word learning and if this maintains one day and one week later. To do this, we conducted three experiments in which adults learned novel animal names in an initial exposure session, followed by three retrieval practice sessions (each containing three tests allowing us to measure orthographic memory) that were either spaced 2 hours apart (spaced schedule) or took place back-to-back immediately after exposure (massed schedule). In addition, in Experiments 1 and 2 (but not Experiment 3), participants completed a follow-up retrieval practice session the day after initial exposure, allowing us to determine whether overnight changes in retention of novel words were affected by the initial learning schedule. All experiments also included a final longer-term retention measure administered one week after initial exposure. Experiment 1 did not include any further exposure to the picture-word pairs after the initial training session (aside from participants seeing the pictures in isolation in the picture naming tests). However, Experiments 2 and 3 included limited re-exposure in the form of corrective feedback incorporated into one of the three retrieval tests (i.e., the base-animal matching task) that was delivered in each retrieval practice session on the first day. The key findings of each experiment are summarised below.

#### *5.1.1.1 Experiment 1*

We did not observe a spacing benefit without feedback on the first day. Instead, we found that adults' abilities to recall the animal names were similar regardless of the schedule condition on Day 1, and their performance was maintained from the first day to Day 2 and one week later. Comparing our findings with other studies that have examined spaced retrieval practice without feedback, the lack of differences between schedules on the first day was surprising. For example, Karpicke & Roediger (2007) and Kroenisen and Kuepper-Tetzel (2021) found a benefit from spaced presentations and within-day spaced sessions without feedback.

A key difference between Experiment 1 and other studies examining retrieval practice without feedback (e.g., Kroenisen & Kuepper-Tetzel, 2021; Karpicke & Roediger, 2007) was that we did not train encoding levels to a pre-set criterion to achieve near-ceiling performance. The main reason for this decision was to actively avoid ceiling level performance, facilitating the detection of potential improvements in new word recall that have been reported in previous studies of word learning over repeated tests in adults (e.g., Kornell, 2009; Roediger & Karpicke, 2006). While we did not see improvements in correct/incorrect recall levels (see Appendix A3), we found small improvements in spelling accuracy across retrieval practice on the first day. This highlights a strength of our experiments: measuring spelling performance rather than correct/incorrect recall. In addition, spelling performance can indicate the quality of orthographic memories (Rossi, Martin-Chang & Ouellette, 2018). Thus, the small but supported improvements from repeated retrieval practice without feedback indicated that repeated retrieval practice could enhance the orthographic quality of the memory representation even without feedback (contrasting Sutterer & Awh, 2016), regardless of which schedule participants were in. Further, we found that repeated retrieval practice prevented a significant decrease in spelling performance across a period of wake, which has been reported (Kurdziel & Spencer, 2016). While we cannot argue that within-day spaced retrieval practice was more beneficial than massed retrieval practice in the absence of feedback, our findings suggest that repeated retrieval practice can enhance the quality of memories for novel words.

Further, we did not observe significant levels of forgetting or improvements overnight or one week later, regardless of whether the initial retrieval schedule was spaced or massed. The lack of spacing benefit again contrast with studies in the spacing literature that report a benefit in long-term measures after spaced retrieval practice without feedback (e.g., Karpicke & Roediger, 2007). It is possible that we did not observe a difference between the schedules because of the relatively low performance, potentially being susceptible to floor effects. On average, participants could only remember between 2-3 full animal names across retrieval practice blocks (Appendix A3). In the spacing literature, a spacing benefit mainly emerges due to less forgetting than after massed retrieval practice. However, because of the low performance after the first day of retrieval practice (2-3 full words recalled at 40-48% matching rate), a possibility is that there was no room for more forgetting in the massed condition, explaining the lack of spacing benefit in Experiment 1. As mentioned above, other studies have countered this by training participants to higher initial performance (between 80 and 100% correct recall, e.g. Kroenisen & Kuepper-Tetzel, 2021; Karpicke & Roediger, 2007). However, this increases the risk of ceiling effects affecting recall. A way to counter this would be to introduce a criterion that participants need to reach by the end of the initial training phase but set the criterion to a level that allows detection of improvements and forgetting in performance (e.g., ~60%).

Overall, the conclusion of Experiment 1 was that retrieval practice was effective in promoting maintained memories of animal names, but counter to our predictions and retrieval difficulty and contextual variability theories, no spacing benefit emerged either on the first, second or final day one week later without feedback or sleep between spaced tests on Day 1, at least when levels of learning are low.

#### *5.1.1.2 Experiment 2*

When feedback was included, the initial learning pattern on Day 1 aligned with extant theoretical accounts (i.e. steeper performance improvements over the first three retrieval practice blocks when they were massed instead of spaced). According to the retrieval difficulty and contextual variability theories, participants in the massed schedule should show steeper improvements in word learning performance because less time has passed between retrieval attempts making retrieval easier (due to less forgetting between attempts), and temporal contextual cues would be similar, aiding initial retrieval. This pattern of initial performance was also found in the second experiment of Karpicke and Roediger (2007), who reported a spacing benefit in long-term measures (10 minutes later).

Interestingly, however, we did not observe a spacing benefit when participants were retested 24 hours and one week later (counter to Karpike & Roediger, 2007). Instead, in the Day 2 retrieval block, the initial benefit from massed retrieval practice disappeared as we saw a steeper overnight improvement for the spaced retrieval practice condition, resulting in a similar performance for both conditions. The effectiveness of overnight consolidation can be affected by the encoding strength of novel memories before sleep (e.g., Denis et al., 2021; Denis et al., 2018; Walker et al., 2019; Tucker & Fishbein, 2008). For example, Denis and colleagues (2018) manipulated the strength of the initial encoding of word pairs by varying the number of exposures, and when tested on recall after a period of sleep, the weakly encoded word pairs showed the least forgetting compared to items with stronger encoding levels. The authors argued that overnight processes prioritised the consolidation of weakly encoded memories. In

Experiment 2, this could have played a role in the steeper rate of improvement observed overnight after spaced retrieval practice, as the levels of encoding were lower than after massed retrieval practice before sleep (as indicated by the lower performance levels). Thus, participants in the spaced schedule could have benefited more from the overnight consolidation than participants in the massed schedule, who mainly strengthened their already strongly encoded memory traces (Petzka et al., 2019).

Alternatively, Bahrick and Hall (2005) highlighted the importance of failure in spaced retrieval attempts. They found that the more failed retrieval attempts that occurred in retrieval practice sessions resulted in improved long-term retention performance (given that feedback was provided). Furthermore, they argued that the failed attempts allowed the novel words to be re-encoded, using the feedback as a mediator. In Experiment 2, we present findings that support this, as adults in the spaced schedules failed more retrieval attempts by the end of the first day but then caught up with the participants practising in massed schedules.

One week later, participants in the three schedules performed at a similar level, meaning that despite the spaced schedule eliciting weaker performance on the first day and steeper overnight improvements (aligning with theoretical accounts), this did not result in a longer-term spacing benefit on Day 7. As discussed in Chapter 2, given the possibility that the Day 2 (post-sleep) retrieval session provided a "spaced" retrieval practice opportunity for both spaced and massed conditions, we reasoned that this "reconsolidation" opportunity may have washed out potential effects of the initial retrieval schedule, masking later schedule effects when memory was measured one week later.

The key take-home message of Experiment 2 was that the retrieval practice and contextual variability theoretical accounts could not explain why we did not see a spacing benefit one week after initial exposure. However, practising retrieval in within-day spaced sessions led to steeper improvements (compared to massed) in recall performance overnight from the first to the second day. This difference between schedules highlights the importance of considering the role of sleep in the spacing effect, as it opens up the possibility that sleep interacts with spacing to elicit a benefit in long-term memory.

#### *5.1.1.3 Experiment 3*

Experiment 3 replicated the same methods as Experiment 2 (i.e., providing feedback in the Day 1 tests) but with the critical exception that no post-sleep retrieval opportunity was provided on Day 2. We again observed an initial massed benefit on the first day, fitting with the retrieval difficulty and contextual variability theoretical accounts. However, unlike our findings of maintained performance in Experiment 2, without a reconsolidation opportunity after sleep, there was more forgetting after massed (than spaced) retrieval practice one week later. Instead, after spaced retrieval practice, performance was maintained at a similar level from the end of Day 1 to Day 7. Importantly, however, performance one week later was similar for both schedules. Therefore, no overall spacing benefit emerged in long-term memory, again counter to the abovementioned theories. As we speculated in Chapter 2, if a final test after a longer period (e.g. one month) was included, the steeper forgetting after massed retrieval practice could have eventually led to a spacing benefit. Nevertheless, we have no evidence here for a spacing *benefit* reported by previous studies one week after initial exposure.

Cross-experiment comparisons of Experiments 2 and 3 confirmed that within-day spaced retrieval practice protected memories from forgetting from Day 1 to Day 7 more than massed retrieval practice. This was potentially due to the harder retrieval (e.g., Pyc & Rawson, 2009) and more varied context (e.g., Pashler et al., 2009) at retrieval attempts, which we argue was reflected in the different performance levels on the first day of testing. Alternatively, decreased forgetting for the spaced condition could also be explained by the weaker levels of encoding at the end of Day 1, as argued above. If weakly encoded memories are prioritised during overnight consolidation (Denis et al., 2020; Denis et al., 2018), and participants practising retrieval in spaced blocks performed lower by the end of Day 1 (indicating weaker encoding levels), overnight consolidation would effectively strengthen these memories, explaining why participants maintained the same performance one week later. In contrast, participants practising massed retrieval already had stronger encoded memory traces, which would not be prioritised during sleep, resulting in a section of animal names not receiving the full benefits from overnight consolidation and being forgotten.

The overall conclusion from Experiment 3 was that within-day spaced retrieval practice protects memories from forgetting, but this effect was not robust enough for a spacing benefit to emerge one week later.

#### *5.1.1.4 Summary of Chapter 2 findings*

In short, the three experiments in Chapter 2 conclude that feedback and sleep are important aspects in the context of spaced versus massed retrieval practice in adults. Once feedback was incorporated with retrieval practice on Day 1, we observed an initial massed benefit that aligned with the retrieval difficulty and contextual variability theories (prior to sleep). However, no spacing benefit emerged one week later, shedding doubt on how well these theoretical accounts can explain the spacing effect on memory retention after longer retention intervals (e.g., one week). In particular, we found that when no spacing benefit emerged one week after initial retrieval practice if no sleep/offline consolidation occurred between initial retrieval attempts, indicating that the retrieval difficulty and contextual variability theoretical accounts provide incomplete explanations without including the caveat of offline consolidation and reconsolidation.

### 5.1.2 Chapter 3: The effect of post-sleep retrieval after within-day spaced and massed retrieval practice in children

From a developmental perspective, Experiments 4 and 5 addressed whether within-day spaced retrieval practice with feedback will result in a spacing benefit immediately, one day, and one week later in samples of school children aged 9-10 years. These experiments adopted the same experimental design as Experiments 2 and 3, with two key differences. First, the experiments took place in primary school classrooms, with whole classes of children. Second, to avoid severe interruptions to the regular school day, we adapted the retrieval practice schedule on the first day to only include two retrieval practice sessions (this will be further discussed in Section 5.3). In addition, to ensure that learning levels remained roughly equal to the adults in Experiments 2-3, we included one additional exposure task in the first session of the first day.

#### *5.1.2.1 Experiment 4*

When children practised retrieval in a classroom setting, we observed continued improvements in memory performance across sessions before and after sleep, with continued improvements between Day 1 and Day 7. These results align with other findings suggesting that children can continue to improve their performance across tests and/or offline consolidation (e.g., James, Gaskell & Henderson, 2019; James et al., 2017; Henderson et al., 2015).

However, these data do not support retrieval difficulty or contextual variability theories for two key reasons. First, we did not find support for a performance difference between the different retrieval schedules on the first day. There was an indication of a trend towards steeper improvements for children in the massed schedule condition (aligning with Experiment 2/3 in adults); however, this did not receive statistical support, potentially linked to the study being underpowered. Second, there was no evidence for spacing benefit on Day 7. Therefore, the main conclusion from Experiment 4 was that children's word learning benefits from repeated retrieval practice regardless of whether the initial retrieval practice is spaced or massed, hence providing preliminary evidence that theories of spaced retrieval practice may not apply to developmental populations.

#### *5.1.2.2 Experiment 5*

A caveat of Experiment 4 was the limited sample size, owing to the Covid-19 pandemic. Experiment 5 addressed this by increasing the sample size, which resulted in greater Bayes Factors indicating stronger effects. In Experiment 5, we also removed the Day 2 follow-up session allowing for direct comparisons with the adult data from Experiment 3, but maintained all other design aspects from Experiment 4.

We found support for an initial benefit from massed retrieval practice, seen as steeper improvements in the word learning tests on the first day. The increased sample size in Experiment 5 (relative to Experiment 4) may have allowed this pattern of massed benefit to emerge as statistically significant on the first day of testing. This contrasts with the typical pattern in the child literature of an early emerging spacing benefit (e.g., Sobel et al., 2011; Childers & Tomassello, 2002). Instead, the benefit from massed retrieval practice on the first day was more in line with the adult-based theoretical accounts, such as retrieval difficulty and contextual variability theories, and mimicking our adult findings of Experiments 2 and 3. This could indicate that these theoretical accounts can account for performance changes in short-term measures in both adults and children, especially before a sleep period.

Interestingly, even in the absence of the post-sleep retrieval opportunity on Day 2, the spaced schedule caught up with those in the massed schedule by Day 7. This resulted in equal performance for all schedule conditions. Thus, initially, Experiment 5 would seem to provide support for the retrieval difficulty and contextual variability theories in child populations. Indeed, the performance changes across the first day align with these theoretical accounts, but the lack of spacing benefit one week later does not. Similar to our adult experiments in Chapter 2, our findings highlight that these theoretical accounts cannot account for all patterns in the spacing literature, especially over longer-term measures.

Focusing on the changes from the first to last day, there was a difference in the trajectory of change for children with different initial retrieval schedules. Children in the massed schedules maintained their already higher performance, while children practising retrieval in the spaced schedules continued to improve their performance, catching up with those in the massed schedule. While this pattern could be attributed to the different spacing schedules, it could also be linked to the different levels of performance by the end of the first day, similar to what we discussed for adults above. As the spaced condition performed at a lower level (indicating weaker encoding strength), they could benefit more from overnight consolidation than those in the massed schedule condition as they performed higher. This benefit could manifest as continued improvements, as children can improve in word learning tasks after overnight sleep (e.g., Henderson et al., 2015; William & Horst, 2014). In contrast, children in the massed schedules also benefited from overnight consolidation, but at a slightly reduced magnitude, manifesting as maintenance of the already higher performance. This highlights that adults and children may show sensitivity to consolidation benefits depending on encoding strength before sleep. This is also in line with the sleep literature where children often show improvements over sleep periods (e.g., James, Gaskell & Henderson, 2019; Ashworth, Hill, Karmiloff-Smith & Dimitriou, 2013; Henderson et al., 2012). However, under conditions of weaker encoding, for example, in populations with dyslexia or neurodevelopmental disorders (e.g., Leonard et al., 2020; Haebig et al., 2019; Smith et al., 2018) these improvements may be reduced.

#### *5.1.2.3 Summary of Chapter 3 findings*

Experiments 4 and 5 found that within-day spaced retrieval practice in a classroom setting did not result in a spacing benefit in children's word learning one week later. Indeed, despite some evidence that spacing initial retrieval practice permits children to catch up to peers who initially undergo massed retrieval practice one week later (potentially signalling that spacing retrieval practice somehow changes the course of offline consolidation), there was no evidence that this led to a long-term advantage here.

### 5.1.3 Chapter 4: Direct comparison of the effects of within-day spaced retrieval practice in adults and children

Finally, in Chapter 4, we compared data from the previous experiments to determine whether adults and children show differences in their word learning trajectories following within-day spaced versus massed retrieval practice. This provides the first direct comparison of within-day spaced retrieval practice in adults and children in a preliminary attempt to address the empirical gap of developmental comparisons in the spacing literature. We conducted two analyses in this chapter: (i) comparing adult and child data from Experiments 2 and 4 to determine whether adults and children benefit from within-day spaced and massed retrieval practice when they also receive a post-sleep retrieval (or "reconsolidation") opportunity, and (ii) comparing adult and child data from Experiments 3 and 5 to examine developmental differences when there was no post-sleep retrieval opportunity. Therefore, our direct comparisons allowed us to determine if developmental differences in word learning following spaced versus massed retrieval practice depend on whether there is a reconsolidation opportunity. This is important because previous literature has reported developmental differences in the effects of offline consolidation on later retention of word learning. For example, James (2019) found larger consolidation effects in children and further suggested that these differences were larger at 7 days than at 24 hours after initial training. However, it remains uncertain whether this was because of repeat testing or an accumulation of overnight consolidation from several consecutive nights' sleep. Our analyses in Chapter 4 therefore allow us to manipulate the presence of an additional test to see whether this enhances a potential developmental difference in the retention of novel words.

#### *5.1.3.1 Analysis 1: comparing the effects of within-day spaced retrieval practice on word learning performance in adults and children before and after sleep and one week later*

On the first day, we found that adults' word learning performance improved at a steeper rate than children's, even when we controlled for the uneven number of retrieval occasions between the experiments (i.e., adults completed three retrieval blocks on the first day and children completed two). This result is in line with literature suggesting that adults can learn several words in a short period and before sleep-based consolidation (e.g., Bahrick & Hall, 2005). In contrast, children improved slower than adults, despite only having to learn 12 novel animal names (relative to adults who learned 20 words). A potential reason for this could be that adults have a vaster knowledge of words in general (see James, Gaskell, Henderson, 2019), meaning that they could link the novel words with prior knowledge to aid learning.

We also found that massed retrieval practice led to steeper improvement rates than spaced retrieval practice, regardless of age. Thus, both adults and children initially follow a pattern of performance that can be explained by retrieval difficulty and contextual variability theories. Specifically, the steeper rates of improvement in recall levels in the massed schedules could have indicated that retrieval was easier (due to less forgetting occurring between retrieval attempts due to less time passing) and more overlapping temporal contextual cues aiding the retrieval (due to the temporal context remaining similar as less time passing between retrieval attempts). This suggests that an initial massed benefit is not developmentally sensitive (at least when considering the age groups under investigation here).

Turning to the comparisons across days, we found that adults and children changed their performance at different rates from Day 1 to Day 2 and Day 7. Specifically, adults only showed improved performance from Day 1 to Day 2 if they were in the spaced condition; otherwise, they maintained their performance, while children showed continued improvements regardless of which retrieval schedule they adhered to. We therefore provide novel empirical evidence of a developmental difference in changes in memory retention overnight and one week later. Specifically, children's improvements in memory retention with repeated tests are more robust than adults' and do not vary as a function of the initial retrieval schedule.

These findings indicate that despite adults and children showing similar word learning benefits from spaced and massed retrieval practice before sleep (i.e., steeper initial learning rates on Day 1) once a reconsolidation opportunity was provided 24 hours later, theoretical accounts (such as retrieval difficulty and contextual variability) cannot explain the observed developmental differences in memory retention one week later. This pattern is consistent with adults and children showing different sleep effects after learning novel words (e.g., adults showing strengthening and maintenance after sleep and children showing continued improvements), highlighting the importance of examining the effects of consolidation/reconsolidation in the spacing effect in adults and children.

##### *5.1.3.2 Analysis 2: comparing the effects of within-day spaced or massed retrieval practice on retention of novel words one week later*

Performance on the first day showed the same as the previous analysis (i.e., Experiment 2 and 4 comparisons), where adults improved faster on the recall tasks than children across the retrieval blocks on the first day, and steeper overall improvements from massed retrieval practice compared to spaced retrieval practice for both age groups. This further showed that within-day spaced or massed retrieval practice initially follows the retrieval difficulty, and contextual variability accounts for the spacing effect before sleep. More importantly here, the analysis of performance changes from Day 1 to Day 7 revealed that both adults and children who practised retrieval in spaced tests on the first day improved their performance one week later, while those who practised massed retrieval maintained their performance from Day 1. Importantly, we did not observe developmental differences depending on the initial retrieval schedule if no retrieval opportunity was provided the day after initial practice. This suggests that children could potentially utilise the post-sleep test to enhance the memory traces of the novel animal names to a greater extent than adults.

Together, these results indicate that within-day spaced retrieval practice protects memories from forgetting the same for both adults and children, but developmental differences emerged once an additional retrieval opportunity was provided after one night's sleep.

## 5.2 Implications

The results of the above experiments suggest that within-day spaced retrieval practice is not enough to elicit a spacing benefit one week later. This contrasts with the broader spacing literature that frequently reports a spacing benefit, particularly in longer-term follow-up tests taking place days, weeks or months after initial learning. Instead, we have shown that a spacing benefit does not emerge in long-term measures if initial retrieval practice was spaced over a single day (i.e., without intervening sleep between initial retrieval attempts) or when feedback was excluded or included.

The discrepancy between our experiments and published studies may be partly due to a publication bias. Spaced retrieval practice has been promoted as a "neuro hit" in educational settings (Surma et al., 2018), meaning that publications would strive to highlight the beneficial aspects of the method and/or that journals would be less likely to accept null results for publication, leading to a skewed perception of the effectiveness of spaced retrieval practice. Publication bias is a problem reaching all research areas and can have adverse effects on research (e.g. meta-analyses are based on an incomplete set of findings) and real-life applications of research findings (Mlinarić, Horvat & Smolčić, 2017; Joober et al., 2012). In the current area of research, it could mean that spaced retrieval practice is used in settings despite lacking the fundamental aspects required for a benefit to emerge. However, before considering the strong conclusion that previous reports of a spacing benefit represent a publication bias, we critically and systematically examine the role of key design aspects (focusing on the roles of sleep and feedback) to better understand why the present results deviate from previous literature and thus reveal the conditions under which a spacing benefit may emerge.

### 5.2.1 The sleep effect in spaced retrieval practice

Sleep is crucial in word learning. As reviewed in Chapter 1, neurological activation during sleep can work to strengthen memories of new words and integrate them with existing knowledge in both adults and children (e.g., Williams & Horst, 2014; Henderson et al., 2012; Tamminen et al., 2010; Davis & Gaskell, 2009). Thus, considering the long-term effects of spaced versus massed retrieval practice in word learning without also considering how sleep might interact with these effects will inevitably result in an incomplete interpretation. For example, Chapters 2 and 3 show that no spacing benefit emerges one week later when retrieval practice is spaced over a day (without intervening sleep). This could suggest that previous reports of a spacing effect (where there was sleep between spaced retrieval practice opportunities) could be due to sleep rather than spacing. However, this needs to be determined directly in future research that compares the effect of spacing with and without intervening sleep (this will be further discussed in Section 5.4).

There are several similarities between the spacing and sleep literature in behavioural measures. For example, word learning studies find reduced forgetting or improvements after sleep (Abel & Bäuml, 2013; Dumay & Gaskell, 2007; Henderson et al., 2012; Lahl, Wispel, Willigens, & Pietrowsky, 2008; Drosopoulos et al., 2007). This pattern is similar to studies comparing overnight spaced sessions (showing reduced forgetting) with a massed session on a single day (showing greater rates of forgetting; e.g., Bell et al., 2014; Karpicke & Bauernschmidt, 2011; Monazedhi et al., 2008; Bloom & Shuell, 1981). Thus, sleep and spacing can both impact the learning and retention of novel words. The question then emerges: which of these effects exerts a greater impact on learning and retention of novel words?

Few studies have attempted to separate the effects of sleep and spaces before the experiments in this thesis. Some exceptions are Kornmeier, Sosic-Vasic and Joos (2022), Bell et al. (2014), and Kroenisen and Kueppler-Tetzel (2021). Bell et al. (2014) found that spaced sessions separated by sleep resulted in better retention of novel words one and four days later compared to massed or spaced sessions not spanning a sleep session. Kroenisen and Kuppler-Tetzel (2021) found that a period of sleep immediately after massed retrieval practice removed the spacing benefit, while massed retrieval practice not followed by a period of sleep resulted in lower performance due to greater levels of forgetting (see Chapter 2 for a full outline of their design). While our experiments did not include an across-sleep spaced control condition, our findings clearly show that when sleep was removed from the interval between spaced retrieval practice opportunities, we did not observe a spacing benefit, either after one night's sleep or one week later (Chapters 2 and 3), regardless of whether feedback was included (Chapter 2) and the age of participants (Chapter 4). The lack of spacing benefit one week after initial exposure potentially implies that spaced retrieval practice may require a period of sleep to elicit a spacing benefit. It would be beneficial to conduct an extended version of the current studies to replicate the current findings and include a control condition with spaced retrieval practice in three sessions separated by 24 hours. This spacing schedule would allow potential offline consolidation and reconsolidation to occur, thus, if a spacing benefit emerges in the one-week retention test, it would confirm that offline consolidation and reconsolidation are key parts of the spacing effect. While our experiments clearly show that no spacing benefit emerges without sleep, this extended study would allow us to draw firmer conclusions on sleep's importance in the spacing effect.

A critical question is whether and how spaced retrieval practice influences the process of offline consolidation and later reconsolidation. Based on the steeper overnight improvement found in adults after practising within-day spaced retrieval practice (Experiment 2), we could hypothesise that adults would show an increase in sleep-based consolidation processes following spaced than massed retrieval practice. One study has already examined the relationship between spaced repetitions and memory consolidation (Vilberg & Davachi, 2013). In this study, participants practised word-object and word-scene pairs in a within-subject design where the massed pairs were restudied after 20 minutes and the spaced pairs after 24 hours. The restudy session took place in an fMRI scanner allowing the authors to examine neural activation and connections between brain regions. Twenty-four hours after the restudy session, participants completed a final memory test. Word-object pairs that were remembered correctly in the final test showed greater connectivity between hippocampal and perirhinal cortex regions in the restudy session, but only if they were spaced. Similarly, forgetting of spaced word-object pairs in the final test could be predicted by the connectivity between the regions, but this could not be made for the massed word-object pairs. Their findings show that when more time and, more importantly, a period of sleep allowing consolidation occurred, the spaced repetition was more effective than the massed repetitions. This provides tentative support for the reconsolidation account for the spacing effect, however, it remains uncertain whether this would be the same in novel word-learning tasks and from spaced retrieval practice rather than restudy.

In sum, taking previous literature and the current experimental findings together, it appears that sleep and spacing effects can sometimes be confused. We have shown that no long-term spacing benefit emerges without intervening sleep between retrieval attempts, and this contrasts with previous studies that incorporate intervening sleep and more consistently report long-term spacing benefits. This could be linked to a publication bias in the literature (where positive results are more likely to be published than negative results). However, we also argue that sleep and its associated offline consolidation and later reconsolidation could have a greater effect than other theoretical accounts of the spacing effect. Subtle evidence of a spacing benefit in the present data (i.e., steeper overnight improvements following spaced versus massed retrieval practice) also points to a role of sleep in influencing memory retention following spaced retrieval practice.

### 5.2.2 Developmental differences in word learning following spaced and massed retrieval practice

A key research question of this thesis is whether adults and children show differences in word learning performance after within-day spaced and massed retrieval practice. Therefore, we allowed for a direct comparison of developmental differences in within-day spaced retrieval practice. There have been studies exploring spaced retrieval practice at different ages of children (e.g., Goossens et al., 2016; Vlach & Sandhofer, 2012) and different ages of adults (e.g., Logan & Balota, 2008), but, as outlined in Chapter 4, we are unaware of any studies directly comparing the two. In light of our analyses and findings in Chapter 4, we will discuss potential aspects that could affect the spacing effect in adults and children in this section.

#### *5.2.2.1 Similarities in the spacing effect between adults and children*

Chapter 4 shows that adults and children improved their word learning performance in the spaced and massed retrieval practice schedules at a similar rate across retrieval blocks on the first day. Specifically, adults and children improved at a steeper rate from massed retrieval than from within-day spaced retrieval practice. As mentioned, this initial pattern of change in word learning performance follows the expected pattern from retrieval difficulty and contextual variability theoretical accounts. Thus, our findings support these theories for both adults and children in short-term measures (i.e. only covering the initial training period). Interestingly, this contrasts with studies finding an initial spacing benefit in children (e.g., Haebig et al., 2021; Childers & Tomasello, 2002) but is in line with adult studies reporting a massed benefit (e.g., Karpicke & Roediger, 2007).

As previously outlined, a key difference between the current experiment and studies that report a spacing benefit in children is that a period of sleep occurs between retrieval attempts (e.g., Goossens et al., 2016; Sobel et al., 2011). However, as mentioned in Chapter 3, the training intensity could also be an underlying reason why some studies report an initial benefit in word learning performance from spaced retrieval practice in children. Childers and Tomasello (2002) showed that children benefit more from shorter and less intense training sessions spaced over multiple days than from longer, more intense sessions on a single day. Thus, it is possible that since we adapted the intensity of learning to be at an appropriate level for both adults and children (adults having to learn 20 words while children learned 12), the learning load was within the children's ability to learn words without the need for offline consolidation. Further, suppose the learning intensity was too easy for adults. In that case, benefits from retrieval practice can be reduced if there was a space between retrieval attempts (Paik & Ritter, 2015), potentially explaining that adults and children benefited from massed retrieval practice on the first day of testing in the current experiments.

The above discussion highlights a potential implication for further discussion in educational settings for both adults and children. Namely, massed retrieval practice can have pedagogical value in classrooms under certain conditions. For example, if the learning load is predicted to be light, teachers could utilise massed retrieval practice to allow fast initial learning of novel words. In contrast, if the learning load is predicted to be high, it might be better to space the practice into shorter sessions that span a period of sleep (based on the extant spacing literature, e.g., Bell et al., 2014; Childers & Tomasello, 2002). However, this remains to be tested directly.

#### *5.2.2.2 Developmental differences in the effects of within-day spaced retrieval practice*

A behavioural difference in behavioural patterns on the first day is that adults improved their performance from the first to final retrieval practice block on day one at a steeper rate than children (regardless of initial retrieval schedule) in both experimental comparisons outlined in Chapter 4. This could be linked to adults having more prior knowledge of vocabulary and semantic knowledge. That is, given the present stimuli were selected to be linkable to prior knowledge (i.e., rare animals that were associated with familiar animals), it is possible that adults could link the novel animal names with prior knowledge to aid recall performance more than children, who would be assumed to have a smaller and shallower vocabulary (Groch et al., 2016; Wilhelm et al., 2012; Wilhelm et al., 2008). Previous studies have shown that both adults and children can use prior semantic knowledge to aid word learning before sleep, but that children benefit more from overnight consolidation, seen as steeper rates of improvements in recall performance (e.g., James, Gaskell & Henderson, 2019; Weighall et al., 2016; Henderson et al., 2015). Further, individual differences in prior knowledge have been found to affect word learning (James et al., 2017). For example, Henderson et al. (2015) found that children aged 7-10 with greater expressive vocabulary showed greater overnight improvements in cued recall performance. This highlights the importance of ensuring that the learning material is appropriately semantically related to prior knowledge when encouraging word learning in adults and children and that individual differences are considered in future studies (see Section 5.4 for further discussion). Thus, underlying developmental differences in prior knowledge could have contributed to the age differences on day 1 in the current experiments.

As outlined previously, no developmental difference emerged in the effects of retrieval schedules on Day 1, but once a period of sleep occurred (allowing potential offline consolidation and reconsolidation), we observed developmental differences in performance changes. Namely, adults who practised spaced retrieval showed similar improvements from Day 1 to Day 2 to children, but not adults who practised massed retrieval. As discussed earlier, overnight improvements for spaced retrieval practice were potentially linked with the lower encoding levels for the spaced (than massed) condition before sleep, allowing offline consolidation to boost reconsolidation on Day 2. However, a key aspect of Smith and Scarf's (2017) argument of reconsolidation in the spacing effect over longer timescales is that reconsolidation processes are more effective after changes in the state of a novel memory. Offline consolidation changes the state of a novel memory by strengthening its representation in neocortical regions (Stickgold & Walker, 2007). Thus, as more effective offline consolidation can occur in children and potentially in our spaced adults, the following reconsolidation processes in these populations could be more effective, seen as the steeper improvements and maintained performance one week later.

### 5.2.3 An adaptive theoretical account of spaced retrieval practice

As discussed in the above sections, our initial behavioural findings (i.e., Day 1) fit with theoretical accounts of contextual variability and retrieval difficulty. However, they cannot account for the changes in behaviour one week later, especially if a period of sleep intervenes in retrieval practice occasions (i.e., the sleep period between Day 1 and Day 2 in Experiments 2 and 4). We have also provided tentative evidence that reconsolidation effects may explain developmental differences in the spacing effect. Here we will collect the key points of our previous discussions and how current theoretical accounts can be adapted based on this.

#### *5.2.3.1 Theoretical adaptation depending on initial encoding levels*

Predicted initial encoding levels must be considered when discussing underlying theoretical explanations of the spacing effect. For example, it is possible that if initial encoding levels are high, retrieval difficulty and contextual variability theories are accurate, as the main purpose of spaced and massed retrieval practice is to slow forgetting. As seen in Karpicke and Roediger (2007), when initial levels of encoding were high (~98% after massed, and 78% after spaced), the increased difficulty and varied contextual cues in spaced presentations successfully slowed forgetting 10 minutes later. Nevertheless, as we have shown, when initial encoding levels are low, and the timespans are longer (i.e., 2 hours between spaced retrieval practice and 24h and one week between retention intervals), reconsolidation effects overshadow these accounts. This is potentially because there is more room for further learning, meaning the beneficial effects of sleep-based consolidation and reconsolidation are more evident than if the initial encoding is close to ceiling effects. As further learning can occur if initial learning exposures are limited, feedback is crucial to spaced retrieval practice after limited initial exposure and should be linked to theoretical explanations of the spacing effect.

If initial learning levels are low (i.e., limited initial exposure to the stimuli), we propose that feedback plays a crucial role in the theoretical underpinnings of the spacing effect. First, we showed that feedback allows for rapid learning if included in massed retrieval practice sessions for adults and children (Experiments 2-5). Supporting this, we did not observe a difference between participants in spaced or massed schedules if no feedback was provided (Experiment 1), indicating that the retrieval difficulty and contextual variability theories cannot be applied to spaced and massed retrieval practice without feedback. Further, when introducing feedback in Experiments 3 and 5, we showed that adults and children could utilise the corrective feedback in the base animal match test to enhance recall in later retrieval attempts, especially if they occurred immediately after (i.e., massed schedules). Thus, feedback made the massed recall attempts easier due to being presented closer in time, and the feedback presentation could be linked to the target memory as a contextual cue, enhancing initial learning levels. However, despite utilising the added feedback immediately, they could not retain the high performance one week later (unless a retrieval attempt occurred after a sleep period; Experiments 2 and 4). Instead, if retrieval practice was spaced 2-3 hours apart, feedback aided initial retrieval attempts to a smaller degree, but it was successfully incorporated with the target memory, aiding later retention one week later (Experiments 3 and 5). This could be attributed to more time passing between retrieval attempts and feedback on Day 1, making retrieval harder and slowing the memory traces' decay according to the retrieval difficulty theories (Pyc & Rawson, 2009).

When initial encoding levels are low, the impact of sleep could be clearer than if initial learning levels are high, especially when feedback is included in initial retrieval attempts. Adults practising retrieval practice in within-day spaced sessions showed that they could retain the feedback from the first day of testing over several consecutive nights’ sleep. As discussed in previous sections, adults could use prior knowledge of words (and familiar animals in the current studies) to learn novel words and incorporate feedback. These connections were then strengthened during overnight sleep, resulting in the maintenance of the knowledge of the novel animal names and feedback (Experiment 3). However, children practising retrieval in within-day spaced sessions continued to improve after several nights of sleep (Experiment 5). This indicates that children could continue to form connections between the target memories and feedback during offline consolidation (Friedrich, Wilhelm, Born & Friederici, 2015). Thus, theoretical accounts of the spacing effect need to include the aspect of sleep-based consolidation and how this process may differ through development. This will be further discussed in the section below.

In short, if initial encoding levels are predicted to be high, the harder retrieval and more varied context from spaced retrieval practice can slow forgetting to later retention measures, compared to easier retrieval and limited contextual variability from massed retrieval. However, if the initial encoding is predicted to be lower and feedback is included in the initial retrieval attempts, the potential effects of sleep-based consolidation and later reconsolidation impact later word retention. Below, we will discuss the potential impact of offline consolidation and reconsolidation on theoretical accounts of spaced retrieval practice.

#### *5.2.3.2 Considering a theoretical distinction between spaced retrieval practice with and without intervening sleep*

Theoretical accounts need to distinguish between spaced retrieval practice with and without intervening sleep. We have shown that if initial retrieval practice takes place within a single day (i.e., without intervening sleep), spaced blocks of retrieval practice protect novel memories from forgetting more than massed blocks in adults (Experiment 3). Despite no spacing benefit emerging in the current experiments, these findings partially fit with the theoretical accounts of retrieval difficulty and contextual variability. A key argument of both these accounts is that spaced retrieval practice slows forgetting more relative to massed retrieval practice through slower decaying activation of memory traces (retrieval difficulty theories; Pyc

& Rawson, 2009) and a greater variety of contextual cues to aid later retrieval (contextual variability theories; Pashler et al., 2009). Thus, our findings suggest that these theoretical accounts can be applied in spaced retrieval practice without intervening sleep in adults. However, these theoretical accounts are not nuanced enough to explain why children showed continued improvements in performance (Experiments 4 and 5) or why adults who practised spaced or massed retrieval maintained performance if a retrieval opportunity occurred after a period of sleep (Experiment 2).

When retrieval practice blocks occurred after a period of sleep, the memory of the novel words was protected from forgetting regardless of the initial retrieval practice schedule in adults (Experiment 2), and children continued to improve their word learning performance (Experiment 4). Thus, the reconsolidation opportunity potentially occurring after a period of offline consolidation effectively masked the initial effects of spaced retrieval practice without an initial intervening sleep period. Similarly, Kroenisen and Kuepper-Tetzel (2021) found that a sleep period overshadowed the initial effects of spaced or massed retrieval practice without intervening sleep. This indicates that if at least one sleep period is predicted to occur between spaced retrieval attempts, the reconsolidation account (Smith & Scarf, 2017) would be the dominant theory. Thus, the reconsolidation account overshadows the retrieval difficulty and contextual variability theories. In short, depending on the planned schedule of retrieval practice, different theoretical accounts are active. Specifically, if a period of sleep occurs between retrieval practice sessions, the effects from sleep and reconsolidation are stronger than effects from retrieval difficulty and contextual variability effects.

As shown through the discussions above, theories of retrieval difficulty and contextual variability can explain certain patterns of the spacing literature but are overshadowed by the effects of a post-sleep retrieval opportunity. Thus, we propose that several theoretical accounts underly the spacing effect and which is dominant depends on the schedule of the retrieval practice. Specifically, if retrieval practice is planned to take place on a single day (like in our experiments), the initial learning performance (i.e., on Day 1) of both adults and children follows the predicted pattern of retrieval difficulty and contextual variability accounts if feedback is included in retrieval attempts (i.e., steeper learning in massed retrieval practice). However, if at least one retrieval practice session takes place after a sleep period, the reconsolidation account would be the dominant theory for performance after a longer retention interval. Thus, developmental differences in sleep-based consolidation effects explain the developmental differences observed in the current experiments (adults showing maintained performance and children showing continued improvements one week after initial learning) and extant literature (e.g., James, 2019; Miles, 2014; Ambridge et al., 2006).

## 5.3 General limitations

We have already mentioned some limitations of the current experiments and how they can be addressed in future studies in the other sections of the General Discussion chapter. Here we will outline the key limitations that need to be addressed to further our understanding of the spacing effect in adults and children.

Several studies have found that repeated retrieval attempts result in better long-term retention of novel words than a single retrieval attempt (e.g., Pyc & Rawson, 2008; Karpicke & Roediger, 2006). However, there have also been reports of diminished benefits from retrieval attempts when a number of successful retrievals have occurred (e.g., Rawson & Dunlosky, 2011). Applying this to spaced and massed retrieval practice, the likelihood of several successful retrievals increases in massed retrieval attempts as a shorter time has passed, potentially reducing the effectiveness of repeated retrieval. In addition, the potential influence of repeated retrieval attempts highlights a limitation of the current experiments. Specifically, participants in Experiments 2 and 4 received one additional retrieval opportunity on Day 2, compared to participants in Experiments 3 and 5 (where the Day 2 session was removed). The reduced number of retrieval attempts explain why participants in Experiments 3 and 5 performed overall lower, but it is also possible that this reduction affected long-term retention. Similarly, adults completed one more retrieval attempt compared to children on Day 1, potentially affecting the results. Thus to be confident in our findings, a further study needs to utilise the same design as the current experiments but ensure that initial retrieval practice is matched in terms of the number of retrieval attempts in both adults and children (i.e., all participants complete the same amount of retrieval practice regardless of practice schedule). Conducting this study would allow us to confidently claim that our findings can be attributed to the experimental manipulation of the retrieval schedule and not affected by the uneven number of retrieval attempts.

Based on our findings and the surrounding literature, we have argued that sleep effects could account for more of the spacing effect than other theoretical accounts. However, this is a speculative argument as we did not have a control condition which spanned spaced retrieval practice sessions across three consecutive days. Alternatively, a nap study could be conducted where spaced retrieval practice sessions occur before and after a daytime nap in a sleep lab. Daytime naps have been found to affect memory similarly to overnight sleep (e.g., van Rijn et al., 2021; Scullin et al., 2017). Thus, a nap period could allow for reconsolidation in a post-nap retrieval block (Smith & Scarf, 2017). Based on this, a hypothesis could be made that a spaced condition that spans three consecutive days or a nap would show better long-term memory performance than within-day spaced or massed retrieval practice due to sleep-based consolidation and reconsolidation. If experimental findings support this hypothesis, we can strongly argue that sleep and reconsolidation deserve a greater space in the theoretical accounts of the spacing effect and should receive more focus in recommendations in educational settings.

Finally, it is unclear how our findings would translate to other learning environments. For example, while we used realistic stimuli, the animal names were niche and unusual but had a strong semantic link to an already known animal. Semantic meaning has been found to benefit word learning (e.g., Takashima et al., 2017), and as outlined above, prior knowledge of semantic memories can affect word learning (e.g., Weighall et al., 2016; Henderson et al., 2015). Thus, the close semantic relationship of our stimuli with prior knowledge could have affected the outcome of our experiments, and it remains unclear if we would see the same results if the learning material did not have a semantic meaning. For example, if the stimuli were animals that did not have an already familiar counterpart, it is possible that we would not observe steeper learning improvements on the first day after massed retrieval practice. Similarly, while it was possible to carry out our retrieval practice paradigm in a realistic learning setting (i.e., a classroom), the actual training tests were explicitly focused on retrieving specific stimuli. Thus, it is not clear whether these findings would be replicated in studies using more implicit learning paradigms, such as storybook learning which is a common exposure through development (e.g., Lokhandwala & Spencer, 2021; Williams & Horst, 2014).

## 5.4 Future directions

Future studies could consider individual differences in prior knowledge and vocabulary carefully. As outlined in Chapter 3, individual differences in vocabulary skills can predict overnight changes in word learning tasks. However, we failed to find a correlation between overnight change and vocabulary IQ in Experiment 4, which was potentially due to the study being underpowered. To explore this potential relationship more carefully, we need to replicate Experiment 4 with a more appropriate sample size.

Another individual difference that could affect the effectiveness of spaced retrieval practice is memory abilities. For example, children with weaker memory abilities may benefit more from massed retrieval practice schedules than spaced ones. For example, according to the contextual variability theoretical accounts, the effectiveness of spaced retrieval practice comes from the additive effect of multiple contextual cues binding with the target memory, aiding later recall through various cues to trigger retrieval of the target memory. Thus, the previous retrieval attempt must exist as memory traces. Children with weaker memory abilities risk forgetting aspects of previous retrieval attempts if the space is too long, failing to add the old and new context to the target memory. Evidence for this can be found in very young children. For example, 16-month-old children, who forget novel information very rapidly, could learn novel words in a massed schedule but not in an interleaved/spaced schedule (Vlach & Johnson, 2013). In contrast, 20-month-old children could learn words in spaced and massed schedules, as they show slower forgetting due to more developed memory mechanisms. Similarly, adults who are more skilled comprehenders could learn more words and show a stronger neurological reaction (stronger N400 ERP at second presentation), indicating more effective word learning than less skilled comprehenders (Perfetti, Wlotko & Hart, 2005). Thus, individual differences in memory ability could also affect the efficiency of spaced retrieval practice in older children and adults. Including measures of general memory ability, such as the revised Wechsler Memory Scale (Elwood, 1991; Wechsler, 1945), in studies exploring spaced retrieval practice would allow for critical correlations to be run to assess the effectiveness of a spaced schedule depending on general memory abilities.

Further, considering individual differences in spaced retrieval practice would develop our understanding of whether spaced retrieval practice is an effective tool in supporting children with developmental learning difficulties. For example, Haebig et al. (2019) and Leonard et al. (2020) have examined spaced and immediate repeated retrieval practice in typically developing (TD) children and children with Developmental Language Disorder (DLD). They found that both TD and DLD children benefited from spaced retrieval practice. Another important area would be to extend our studies to populations with dyslexia. For example, children with dyslexia have been found to show lower levels of long-term retention of novel words than TD children (e.g. Menghini et al., 2010). This could be due to a range of cognitive differences, such as encoding difficulties when learning novel words. Additionally, it has been found that the number of sleep spindles did not correlate with overnight changes in word learning performance in children with dyslexia (Smith et al., 2018), so further examining learning methods in dyslexic populations is important. Finally, spaced retrieval practice has been suggested to reduce the cognitive learning load due to more time and potential intervening consolidation; thus, it can be hypothesised that children with dyslexia could show greater spacing benefits than TD children.

## 5.5 Final conclusions

Throughout this thesis, we have examined methods for enhancing word learning and retention in adults and children. We utilised the popular method of spaced retrieval practice but distinguished between the spacing effect and potential sleep effects by removing intervening sleep periods between test sessions. Across five experiments, we have shown that spaced retrieval practice will not always result in better long-term retention of novel words than massed retrieval practice.

First, we found that adults and children learn novel animal names more effectively if they practised retrieval in a massed session that included feedback (Chapters 2 and 3). This was in line with popular theories of retrieval difficulty and contextual variability. According to these theories, we expect better animal name retention after spaced retrieval practice. While we observed tentative evidence that within-day spacing may protect against forgetting one week after initial encoding in adults (Chapter 2) and allowed continued improvements in children (Chapter 3), we did not observe a spacing benefit. We argue that the initial massed benefit in adults and children follows the expected pattern of retrieval difficulty and contextual variability theories, but these fail to explain the lack of spacing benefit one week later. The easier retrieval and greater overlap of contextual cues resulted in steeper learning rates for adults and children in the massed retrieval practice schedules. If a retrieval opportunity was given after a sleep period, overnight consolidation and reconsolidation improved performance for children in all schedules but only adults who practised retrieval in the spaced schedule. This indicates that adults practising retrieval in a within-day spaced schedule and children, regardless of initial retrieval schedule, relied on reconsolidation for further improvements in word learning performance. Importantly, this highlights that reconsolidation has an impact on the retention of novel words, and studies exploring the spacing effect need to take this into account.

Second, we directly compared adult and child data to compare the effects of within-day spaced or massed retrieval practice (Chapter 4), filling an empirical gap. Interestingly, we did not observe a developmental difference in the effects of spaced retrieval practice on the first day of practice, but we observed a developmental difference in the effects of spaced retrieval practice overnight and one week later. Therefore, we provide evidence that adults and children benefit from spaced and massed retrieval practice differently if a retrieval session occurs after a period of consolidation. This could be attributed to developmental differences in the effectiveness of offline consolidation and subsequent reconsolidation, where children have been found to show more effective offline consolidation than adults. We argue that the retrieval difficulty and contextual variability theories can explain the initial word learning performance from within-day spaced and massed retrieval practice in adults and children. But once a period of sleep and subsequent reconsolidation occurs, the reconsolidation theory outweighs the previous effects, and due to underlying developmental differences in sleep-based consolidation processes, we observed developmental differences in memory retention one week later. This highlights the importance of continuing to research spaced retrieval practice in adults and children and how overnight consolidation and reconsolidation affect the long-term retention of novel words.

To conclude, this thesis shows that the spacing effect observed in the literature is not guaranteed and can be affected by aspects such as feedback and reconsolidation opportunities. Further, we have shown that massed retrieval practice can be more effective in short-term word learning and equally as effective as within-day spaced retrieval practice when measured one week later. Finally, we propose that retrieval difficulty and contextual variability theories can account for the initial learning behaviour of adults and children, but the effects from a period of offline consolidation and subsequent reconsolidation overshadow these theories if a period of sleep occurs between retrieval attempts. This shift in the dominant underlying theory of the spacing effect explains developmental differences we observed one week after initial spaced or massed retrieval practice. Finally, this thesis shows that the underlying theoretical accounts of the spacing effect depend on several variables, such as feedback and intervening sleep between initial retrieval attempts. We need to examine these closer to determine when a spacing benefit emerges and when it does not.

# Appendices

## Appendix A: Supplementary material to Chapter 2

### *A1. Lists of stimuli used in Experiments 1-3.*

**Experiment 1 stimuli**

| **List 1** | **List 2** | **Base Animal** |
|---|---|---|
| GUNDI | PACA | RAT |
| HOOPOE | TURACO | BIRD |
| NUMBAT | COATI | SQUIRREL |
| CARACAL | MANUL | CAT |
| LANGUR | SIAMANG | MONKEY |
| PECCARY | MEISHAN | PIG |
| OKAPI | QUAGGA | ZEBRA |
| SAIGA | KUDU | DEER |
| BAGOT | ARGALI | GOAT |
| VICUNA | GERENUK | ANTELOPE |
| GLAUCUS | SCOTOPLANE | FISH |
| AGAMA | NOROPS | LIZARD |

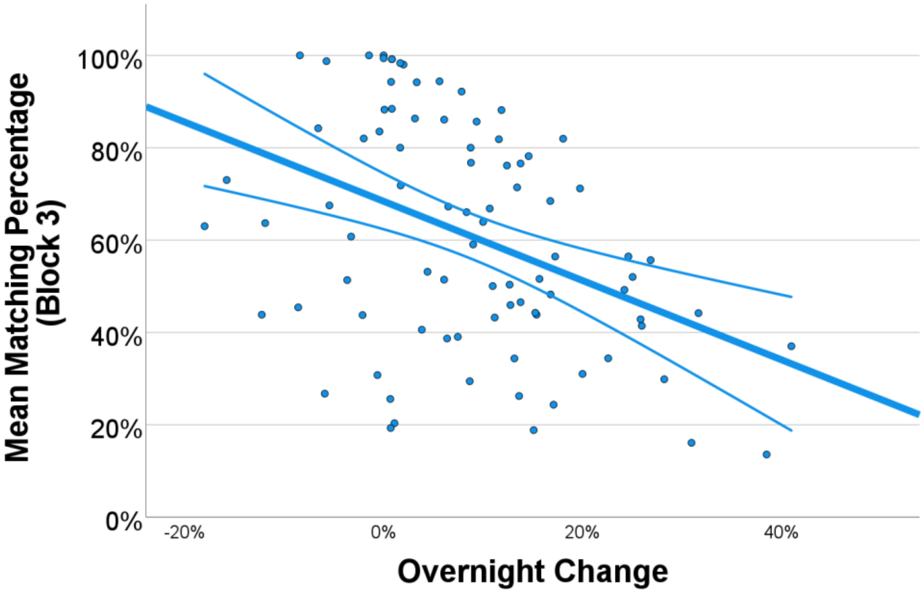
**Experiments 2 and 3 stimuli**

| **List 1** | **List 2** | **Base Animal** |
|---|---|---|
| UMBONIA | MORPHO | BUTTERFLY |
| CIVET | KODKOD | CAT |
| KIRKII | BROCKET | DEER |
| WARRAH | LYCAON | DOG |
| STENELLA | VEXILLIFER | DOLPHIN |
| ISISTIUS | OWSTONI | SHARK |
| FENNEC | DHOLE | FOX |
| BAGOT | URIAL | GOAT |
| JERBOA | SENGI | HAMSTER |
| ECHIDNA | TENREC | HEDGEHOG |
| GLAUCUS | PELAGIA | JELLYFISH |
| AGAMA | NOROPS | LIZARD |
| VICUNA | WANAKU | LLAMA |
| LANGUR | INDRI | MONKEY |
| TURACO | QUETZAL | PARROT |
| PECCARY | XUYEN | PIG |
| XERUS | RATUFA | SQUIRREL |
| OCULIFER | ANGONOKA | TORTOISE |
| NEBLINA | ZORELLI | WEASEL |
| QUAGGA | GREVYI | ZEBRA |

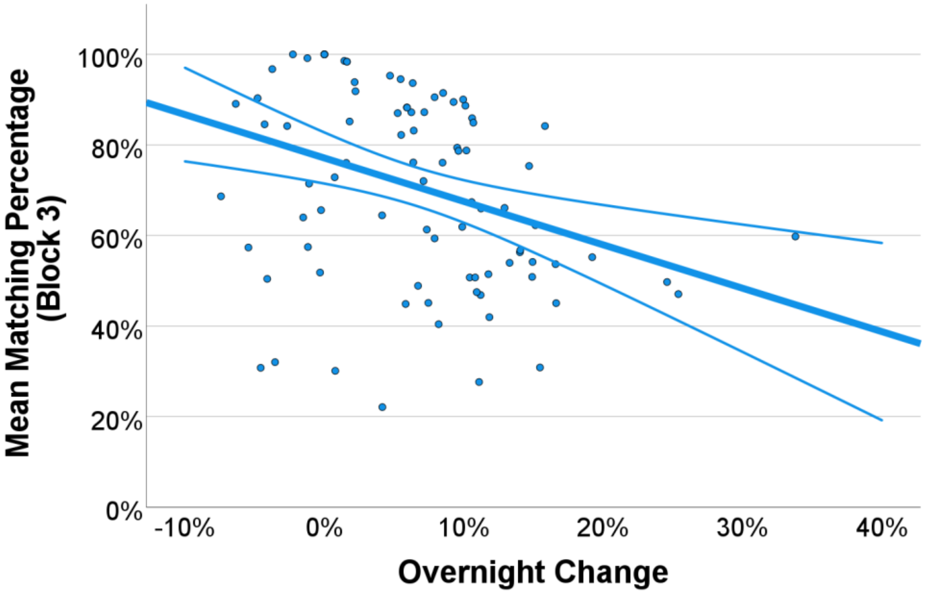
### *A2. Correlation analysis of final performance on Day 1 and overnight change in Experiment 2*

Analysis of correlation between final learning performance on Day 1 and overnight change. To determine whether the overnight change (i.e. B4-B3) in performance was related to the levels of learning by the end of Day 1 (i.e. B3), we ran a Bayesian Kendall's Tau correlation on those variables. There was overwhelming support for a moderate to strong negative correlation in all three tests (cued recall tau = -0.29, $BF_{10}$ = 332.09; picture naming tau = -0.30, $BF_{10}$ = 540.24; base animal match tau = -0.35, $BF_{10}$ = 11172.49). These results suggest that lower performance by the end of Day 1 correlated with greater change overnight (as seen in Figure A2).

**A2.a Cued recall ($R^2$ linear = 0.165)**



**A2.b Picture naming ($R^2$ linear = 0.114)**



**A2.c Base animal match ($R^2$ linear = 0.212)**

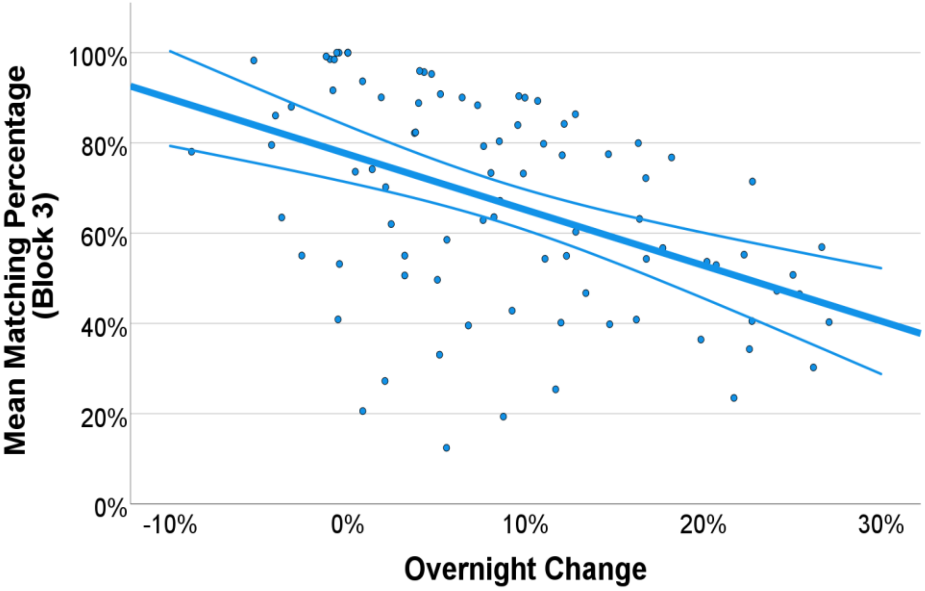


**Figure A2.** Graphs displaying correlation analyses of mean performance in Block 3 and overnight change in performance in the Cued Recall (**A2.a**), Picture Naming (**A2.b**), and Base Animal Match (**A2.c**) tests.

***A3. Mean number of 100% correct recall attempts (± SD) out of 12 animal names in Experiments 1-3.***

| | | DAY 1 | | DAY 2 | DAY 7 |
|---|---|---|---|---|---|
| | **B1** | **B2** | **B3** | **B4** | **B5** |
| **CUED RECALL** | | | | | |
| **Experiment 1** | | | | | |
| Massed AM | 3.0 (± 2.0) | 2.9 (± 2.0) | 3.2 (± 1.8) | 3.3 (± 2.1) | 3.4 (± 2.2) |
| Massed PM | 2.6 (± 1.9) | 2.5 (± 1.9) | 2.6 (± 1.8) | 2.9 (± 1.7) | 2.6 (± 1.9) |
| Spaced | 2.3 (± 1.8) | 2.3 (± 1.4) | 2.2 (± 1.4) | 2.5 (± 1.7) | 2.4 (± 1.5) |
| **Experiment 2** | | | | | |
| Massed AM | 3.8 (± 3.2) | 9.5 (± 4.9) | 13.0 (± 4.9) | 14.0 (± 4.6) | 13.7 (± 4.5) |
| Massed PM | 3.4 (± 3.6) | 7.0 (± 4.9) | 11.2 (± 6.1) | 12.3 (± 6.2) | 12.1 (± 6.3) |
| Spaced | 3.0 (± 1.9) | 4.7 (± 3.0) | 6.4 (± 4.5) | 10.4 (± 4.4) | 10.6 (± 4.7) |
| **Experiment 3** | | | | | |
| Massed | 4.0 (± 3.7) | 6.7 (± 4.9) | 10.1 (± 5.9) | | 7.3 (± 5.2) |
| Spaced | 4.9 (± 3.7) | 6.1 (± 4.3) | 7.7 (± 4.6) | | 8.0 (± 5.1) |
| | | | | | |
| **PICTURE NAMING** | | | | | |
| **Experiment 1** | | | | | |
| Massed AM | 2.1 (± 1.9) | 2.0 (± 1.7) | 2.3 (± 1.7) | 2.4 (± 1.7) | 2.1 (± 1.7) |
| Massed PM | 2.1 (± 1.7) | 2.3 (± 1.7) | 2.2 (± 1.7) | 2.4 (± 1.8) | 2.4 (± 1.6) |
| Spaced | 2.0 (± 2.0) | 2.2 (± 1.4) | 2.1 (± 1.5) | 2.1 (± 1.7) | 2.2 (± 1.7) |
| **Experiment 2** | | | | | |
| Massed AM | 3.8 (± 3.0) | 9.8 (± 5.1) | 13.5 (± 4.2) | 14.7 (± 4.4) | 14.2 (± 4.7) |
| Massed PM | 3.5 (± 4.2) | 7.1 (± 5.5) | 11.5 (± 6.0) | 12.8 (± 5.8) | 12.5 (± 6.1) |
| Spaced | 2.9 (± 1.9) | 4.9 (± 2.9) | 7.3 (± 3.8) | 11.4 (± 4.6) | 10.8 (± 4.3) |
| **Experiment 3** | | | | | |
| Massed | 4.1 (± 4.4) | 7.0 (± 4.9) | 10.8 (± 5.4) | | 7.8 (± 5.3) |
| Spaced | 4.2 (± 3.7) | 5.7 (± 4.4) | 7.6 (± 4.7) | | 8.2 (± 5.2) |
| | | | | | |
| **BASE ANIMAL MATCH** | | | | | |
| **Experiment 1** | | | | | |
| Massed AM | 2.1 (± 1.9) | 2.1 (± 1.8) | 2.5 (± 1.8) | 2.5 (± 2.0) | 2.3 (± 1.7) |
| Massed PM | 1.9 (± 1.5) | 2.1 (± 1.6) | 2.1 (± 1.7) | 2.4 (± 1.7) | 2.1 (± 1.5) |
| Spaced | 1.7 (± 1.3) | 1.7 (± 1.2) | 1.8 (± 1.6) | 2 (± 1.6) | 2.1 (± 1.6) |
| **Experiment 2** | | | | | |
| Massed AM | 3.6 (± 3.0) | 9.2 (± 5.3) | 13.3 (± 4.5) | 14.5 (± 4.4) | 14.5 (± 4.7) |
| Massed PM | 3.2 (± 3.6) | 7.1 (± 5.7) | 11.2 (± 6.2) | 12.5 (± 6.1) | 12.4 (± 6.1) |
| Spaced | 2.3 (± 1.9) | 4.7 (± 2.9) | 6.9 (± 3.8) | 11.4 (± 4.4) | 10.8 (± 4.2) |
| **Experiment 3** | | | | | |
| Massed | 3.9 (± 3.9) | 6.9 (± 5.0) | 10.3 (± 5.5) | | 7.8 (± 5.5) |
| Spaced | 3.8 (± 3.6) | 5.6 (± 4.1) | 7.5 (± 4.6) | | 7.9 (± 5.1) |

## Appendix B: Supplementary material to Chapter 3

### *B1. List of stimuli used in Experiments 4-5.*

**Experiments 4 and 5 stimuli**

| **Novel animal** | **Base Animal** |
|---|---|
| UMBONIA | BUTTERFLY |
| KODKOD | CAT |
| LYCAON | DOG |
| BAGOT | GOAT |
| JERBOA | HAMSTER |
| GLAUCUS | JELLYFISH |
| AGAMA | LIZARD |
| INDRI | MONKEY |
| TURACO | PARROT |
| PECCARY | PIG |
| RATUFA | SQUIRREL |
| QUAGGA | ZEBRA |

# References

Abel, M., & Bäuml, K. T. (2020). Would you like to learn more? Retrieval practice plus feedback can increase motivation to keep on studying. *Cognition*, *201*, 104316. https://doi.org/10.1016/j.cognition.2020.104316

Abel, M., & Bäuml, K. T. (2013) Sleep can reduce proactive interference, *Memory*, *22:4*, 332-339, DOI: 10.1080/09658211.2013.785570

Abel, M., Haller, V., Köck, H., Pötschke, S., Heib, D., Schabus, M., & Bäuml, K. T. (2019). Sleep reduces the testing effect-But not after corrective feedback and prolonged retention interval. *Journal of experimental psychology. Learning, memory, and cognition*, *45*(2), 272–287. https://doi.org/10.1037/xlm0000576

Abott, E. E. (1909). On the analysis of the factor of recall in the learning process. *The Psychological Review: Monograph Supplements, 11*(1), 159–177. https://doi.org/10.1037/h0093018

Aczel, B., Hoekstra, R., Gelman, A., Wagenmakers, E. J., Klugkist, I. G., Rouder, J. N., Vandekerckhove, J., Lee, M. D., Morey, R. D., Vanpaemel, W., Dienes, Z., & van Ravenzwaaij, D. (2020). Discussion points for Bayesian inference. *Nature Human Behaviour 2020 4:6*, *4*(6), 561–563. https://doi.org/10.1038/s41562-019-0807-z

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659-701.

Agarwal, P. K., Bain, P. M., Chamberlain, R. W., Agarwal, P. K., Bain, P. M., & Chamberlain, R. W. (2009). The Value of Applied Research: Retrieval Practice Improves Classroom Learning and Recommendations from a Teacher, a Principal, and a Scientist. *Journal of*

*Experimental Psychology: Applied*, *103*, 437–448. https://doi.org/10.1007/s10648-012-9210-2

Agarwal, P. K., Bain, P. M., & Chamberlain, R. W. (2012). The value of applied research: Retrieval practice improves classroom learning and recommendations from a teacher, a principal, and a scientist. *Educational Psychology Review*, *24*(3), 437–448. https://doi.org/10.1007/s10648-012-9210-2

Agarwal, P. K., D’antonio, L., Roediger Iii, H. L., Mcdermott, K. B., & Mcdaniel, M. A. (2014). Classroom-based programs of retrieval practice reduce middle school and high school students’ test anxiety. *Journal of Applied Research in Memory and Cognition*, *3*, 131–139. https://doi.org/10.1016/j.jarmac.2014.07.002

Alberini, C. M. (2011). The role of reconsolidation and the dynamic process of long-term memory formation and storage. *Frontiers in Behavioral Neuroscience*, *5*, 12. https://doi.org/10.3389/fnbeh.2011.00012

Ambridge, B., Theakston, A. L., Lieven, E. V. M., & Tomasello, M. (2006). The distributed learning effect for children’s acquisition of an abstract syntactic construction. *Cognitive Development*, *21*(2), 174–193. https://doi.org/10.1016/j.cogdev.2005.09.003

Antony, J. W., Ferreira, C. S., Norman, K. A., & Wimber, M. (2017). Retrieval as a Fast Route to Memory Consolidation. *Trends in Cognitive Sciences*, *21*(8), 573–576. https://doi.org/10.1016/j.tics.2017.05.001

Armstrong, R., Arnott, W., Copland, D.A., McMahon, K., Khan, A., Najman, J.M. and Scott, J.G. (2017), Change in receptive vocabulary from childhood to adulthood: associated mental health, education and employment outcomes. *International Journal of Language & Communication Disorders*, *52*: 561-572. https://doi.org/10.1111/1460-6984.12301

Ashworth, A., Hill, C. M., Karmiloff-Smith, A., & Dimitriou, D. (2017). A cross-syndrome study of the differential effects of sleep on declarative memory consolidation in children with neurodevelopmental disorders. *Developmental Science*, *20*(2), e12383. https://doi.org/10.1111/desc.12383

Bahrick, H. P., & Hall, L. K. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language*, *52*(4), 566–577. https://doi.org/10.1016/J.JML.2005.01.012

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014

Barcroft, J. (2007). Effects of Opportunities for Word Retrieval During Second Language Vocabulary Learning. *Language Learning*, *57*(1), 35–56. https://doi.org/10.1111/j.1467-9922.2007.00398.x

Bell, M. C., Kawadri, N., Simone, P. M., & Wiseheart, M. (2014). Long-term memory, sleep, and the spacing effect. *Memory*, *22*(3), 276–283. https://doi.org/10.1080/09658211.2013.778294

Berry, R. B., Wagner, M. H., Berry, R. B., & Wagner, M. H. (2015). Introduction. *Sleep Medicine Pearls*, 10–14. https://doi.org/10.1016/B978-1-4557-7051-9.00002-4

Biemiller, A. (2003). Vocabulary: Needed if more children are to read well. *Reading Psychology*, *24(3-4)*, 323–335. https://doi.org/10.1080/02702710390227297

Bird, S. (2011). Effects of distributed practice on the acquisition of second language English syntax—ERRATUM. *Applied Psycholinguistics*, *32*(2), 435-452. doi:10.1017/S0142716410000470

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the real world: Essays illustrating fundamental contributions to society*, *2*(59-68).

Bloom, K. C., & Shuell, T. J. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *Journal of Educational Research*, *74*(4), 245–248. https://doi.org/10.1080/00220671.1981.10885317

Born, J. (2010). Slow-wave sleep and the consolidation of long-term memory. *The World Journal of Biological Psychiatry*, *11*(sup1), 16-21.

Brown, H., Weighall, A., Henderson, L. M., & Gareth Gaskell, M. (2012). Enhanced recognition and recall of new words in 7- and 12-year-olds following a period of offline consolidation. *Journal of Experimental Child Psychology*, *112*(1), 56–72. https://doi.org/10.1016/j.jecp.2011.11.010

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied*, *13*(4), 273–281. https://doi.org/10.1037/1076-898X.13.4.273

Bäuml, K. H., Holterman, C., & Abel, M. (2014). Sleep can reduce the testing effect: it enhances recall of restudied items but can leave recall of retrieved items unaffected. *Journal of experimental psychology. Learning, memory, and cognition*, *40*(6), 1568–1581. https://doi.org/10.1037/xlm0000025

Cairney, S. A., Durrant, S. J., Jackson, R., & Lewis, P. A. (2014). Sleep spindles provide indirect support to the consolidation of emotional encoding contexts. *Neuropsychologia*, *63*, 285–292. https://doi.org/10.1016/j.neuropsychologia.2014.09.016

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1563–1569. https://doi.org/10.1037/a0017021

Carpenter, S. K., & Agarwal, P. K. (2020). *How To Use Spaced Retrieval Practice To Boost Learning*. Retrieved January 16, 2019, from http://pdf.retrievalpractice.org/SpacingGuide.pdf

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H. K., Pashler, H., Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H. K., & Pashler, H. (2012). Using Spacing to Enhance Diverse Forms of Learning: Review of Recent Research and Implications for Instruction. *Educ Psychol Rev*, *24,* 369–378. https://doi.org/10.1007/s10648-012-9205-z

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology*, *19*(5), 619–636. https://doi.org/10.1002/acp.1101

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & cognition*, *36*(2), 438–448. https://doi.org/10.3758/mc.36.2.438

Carpenter, S. K., & Vul, E. (2011). Delaying feedback by three seconds benefits retention of face-name pairs: the role of active anticipatory processing. *Memory & cognition*, *39*(7), 1211–1221. https://doi.org/10.3758/s13421-011-0092-1

Carskadon, M. A., Acebo, C., & Jenni, O. G. (2004). Regulation of adolescent sleep: implications for behavior. *Annals of the New York Academy of Sciences, 1021*, 276–291. https://doi.org/10.1196/annals.1308.032

Cellini, N., Torre, J., Stegagno, L., & Sarlo, M. (2016). Sleep before and after learning promotes the consolidation of both neutral and emotional information regardless of REM presence. *Neurobiology of Learning and Memory*, *133*, 136–144. https://doi.org/10.1016/J.NLM.2016.06.015

Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2009). Optimizing Distributed Practice: Theoretical Analysis and Practical Implications. *Experimental Psychology*, *56*, 236–246. https://doi.org/10.1027/1618-3169.56.4.236

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. https://doi.org/10.1037/0033-2909.132.3.354

Childers, J. B., & Tomasello, M. (2002). Two-year-olds learn novel nouns, verbs, and conventional actions from massed or distributed exposures. *Developmental Psychology*, *38*(6), 967–978. https://doi.org/10.1037/0012-1649.38.6.967

Cirelli, C., & Tononi, G. (2019). Linking the need to sleep with synaptic function. *Science*, *366*(6462), 189–190. https://doi.org/10.1126/SCIENCE.AAY5304

Coane, J. H. (2013). Retrieval practice and elaborative encoding benefit memory in younger and older adults. *Journal of Applied Research in Memory and Cognition*, *2*(2), 95–100. https://doi.org/10.1016/j.jarmac.2013.04.001

Coccoz, V., Maldonado, H., & Delorenzi, A. (2011). The enhancement of reconsolidation with a naturalistic mild stressor improves the expression of a declarative memory in humans. *Neuroscience*, *185*, 61–72. https://doi.org/10.1016/j.neuroscience.2011.04.023

Cowan, N., & Alloway, T. (2009). Development of working memory in childhood. In M. L. Courage & N. Cowan (Eds.), *The development of memory in infancy and childhood* (pp. 303–342). Psychology Press.

Craik, F. I. M., & Bialystok, E. (2006). Cognition through the lifespan: mechanisms of change. *Trends in Cognitive Sciences*, *10*(3), 131–138. https://doi.org/10.1016/J.TICS.2006.01.007

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*, *14*(3), 215–235. https://doi.org/10.1002/(SICI)1099-0720(200005/06)14:3<215::AID-ACP640>3.0.CO;2-1

Davis, M. H., & Gaskell, M. G. (2009). A complementary systems account of word learning: neural and behavioural evidence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1536). http://rstb.royalsocietypublishing.org/content/364/1536/3773

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). Spacing and Testing Effects: A Deeply Critical, Lengthy, and At Times Discursive Review of the Literature. *Psychology of Learning and Motivation - Advances in Research and Theory*, *53*(C), 63–147. https://doi.org/10.1016/S0079-7421(10)53003-2

Demanuele, C., Bartsch, U., Baran, B., Khan, S., Vangel, M. G., Cox, R., Hämäläinen, M., Jones, M. W., Stickgold, R., & Manoach, D. S. (2017). Coordination of Slow Waves With Sleep Spindles Predicts Sleep-Dependent Memory Consolidation in Schizophrenia. *Sleep*, *40*(1). https://doi.org/10.1093/SLEEP/ZSW013

Denis, D., Bursal, V., Oquin, S., Morgan, A., & Stickgold, R. (2018). 0106 The Role Of Encoding Strength In The Prioritization Of Memories For Consolidation During Sleep. *Sleep*, *41*(suppl_1), A42–A42. https://doi.org/10.1093/SLEEP/ZSY061.105

Denis, D., Mylonas, D., Poskanzer, C., Bursal, V., Payne, J. D., & Stickgold, R. (2021). Sleep Spindles Preferentially Consolidate Weakly Encoded Memories. *Journal of Neuroscience, 41*(18), 4088–4099. https://doi.org/10.1523/JNEUROSCI.0818-20.2021

Diekelmann, S., Wilhelm, I., & Born, J. (2009). The whats and whens of sleep-dependent memory consolidation. *Sleep Medicine Reviews*, *13*(5), 309–321. https://doi.org/10.1016/J.SMRV.2008.08.002

Dolan, R., Fletcher, P. Dissociating prefrontal and hippocampal function in episodic memory encoding. *Nature* 388, 582–585 (1997). https://doi.org/10.1038/41561

Drosopoulos, S., Schulze, C., Fischer, S., & Born, J. (2007). Sleep's function in the spontaneous recovery and consolidation of memories. *Journal of experimental psychology. General*, *136*(2), 169–183. https://doi.org/10.1037/0096-3445.136.2.169

Dumay, N., & Gaskell, M. G. (2007). Sleep-Associated Changes in the Mental Representation of Spoken Words. *Psychological Science*, *18*(1), 35–39. https://doi.org/10.1111/j.1467-9280.2007.01845.x

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, Supplement*, *14*(1), 4–58. https://doi.org/10.1177/1529100612453266

Elliott, C. D., Smith, P., & McCulloch, K. (1997). *Technical manual British Ability Scales II*. Windsor, Berkshire: NFERNELSON Publishing Company.

Elwood, R. W. (1991). The Wechsler Memory Scale—Revised: Psychometric characteristics and clinical application. *Neuropsychology Review 1991 2:2*, *2*(2), 179–201. https://doi.org/10.1007/BF01109053

Estes W. K. (1955). Statistical theory of distributional phenomena in learning. Psychological review, 62(5), 369–377. https://doi.org/10.1037/h0046888

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods 2007 39:2*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental science, 16(2*), 234–248. https://doi.org/10.1111/desc.12019

Fletcher, F. E., Knowland, V., Walker, S., Gaskell, M. G., Norbury, C., & Henderson, L. M. (2020). Atypicalities in sleep and semantic consolidation in autism. *Developmental Science*, *23*(3). https://doi.org/10.1111/DESC.12906

Fodor, J. A., Fodor, J. (1983). The Modularity of Mind. United Kingdom: A BRADFORD BOOK.

Forcato, C., Burgos, V. L., Argibay, P. F., Molina, V. A., Pedreira, M. E., & Maldonado, H. (2007). Reconsolidation of declarative memory in humans. *Learning & Memory*, *14*(4), 295. https://doi.org/10.1101/LM.486107

Friedrich, M., Wilhelm, I., Born, J., & Friederici, A. D. (2015). Generalization of word meanings during infant sleep. *Nature Communications 2015 6:1*, *6*(1), 1–9. https://doi.org/10.1038/ncomms7004

Fritz, C. O., Morris, P. E., Nolan, D., & Singleton, J. (2007). Expanding retrieval practice: An effective aid to preschool children's learning. *Quarterly Journal of Experimental Psychology*, *60*(7), 991–1004. https://doi.org/10.1080/17470210600823595

Gais, S., Lucas, B., & Born, J. (2006). Sleep after learning aids memory recall. *Learning & Memory (Cold Spring Harbor, N.Y.)*, *13*(3), 259–262. https://doi.org/10.1101/lm.132106

Gais, S., Mölle, M., Helms, K., & Born, J. (2002). Learning-Dependent Increases in Sleep Spindle Density. *Journal of Neuroscience*, *22*(15), 6830–6834. https://doi.org/10.1523/JNEUROSCI.22-15-06830.2002

Gaskell, M. G., & Dumay, N. (2003). Lexical competition and the acquisition of novel words. *Cognition*, *89*(2), 105–132. https://doi.org/10.1016/S0010-0277(03)00070-2

Gates, A.I. (1917). Recitation as a factor in memorizing. *Archives of Psychology, 40,* Pp. 104.

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, DOI: 10.1080/09658211.2014.944916

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., Tabbers, H. K., Bouwmeester, S., & Zwaan, R. A. (2016). Distributed Practice and Retrieval Practice in Primary School Vocabulary Learning: A Multi-classroom Study. *Applied Cognitive Psychology*, *30*(5), 700–712. https://doi.org/10.1002/ACP.3245

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., Tabbers, H. K., & Zwaan, R. A. (2012). Spreading the words: A spacing effect in vocabulary learning. *Journal of Cognitive Psychology*, 24, 965–971. DOI:10.1080/20445911.2012.722617.

Goossens, N. A. M. C., Camp, G., Verkoeijen, P. P. J. L., Tabbers, H. K., & Zwaan, R. A. (2014). The benefit of retrieval practice over elaborative restudy in primary school vocabulary learning. *Journal of Applied Research in Memory and Cognition*, *3*(3), 177–182. https://doi.org/10.1016/J.JARMAC.2014.05.003

Guo L., (2021) Effects of the initial test interval and feedback timing on L2 vocabulary retention, *The Language Learning Journal*, 49:3, 382-398, DOI: 10.1080/09571736.2018.1551416

Green, J. L., Weston, T., Wiseheart, M., & Rosenbaum, R. S. (2014). Long-term spacing effect benefits in developmental amnesia: Case experiments in rehabilitation. *Neuropsychology*. https://doi.org/10.1037/neu0000070

Groch, S., Schreiner, T., Rasch, B. *et al.* Prior knowledge is essential for the beneficial effect of targeted memory reactivation during sleep. *Sci Rep* 7, 39763 (2017). https://doi.org/10.1038/srep39763

Grosvenor, A., & Lack, L. C. (1984). The Effect of Sleep Before or After Learning on Memory. *Sleep*, *7*(2), 155–167. https://academic.oup.com/sleep/article/7/2/155/2753345

Haebig, E., Leonard, L. B., Deevy, P., Karpicke, J., Christ, S. L., Usler, E., Kueser, J. B., Souto, S., Krok, W., & Weber, C. (2019). Retrieval-Based Word Learning in Young Typically Developing Children and Children With Development Language Disorder II: A Comparison of Retrieval Schedules. *Journal of Speech, Language, and Hearing Research*, *62*(4), 944–964. https://doi.org/10.1044/2018_JSLHR-L-18-0071

Henderson, L., Devine, K., Weighall, A., & Gaskell, G. (2015). When the daffodat flew to the intergalactic zoo: Off-line consolidation is critical for word learning from stories. *Developmental Psychology*, *51*(3), 406–417. https://doi.org/10.1037/a0038786

Henderson, L. M., & James, E. (2018). Consolidating new words from repetitive versus multiple stories: Prior knowledge matters. *Journal of Experimental Child Psychology*, *166*, 465–484. https://doi.org/10.1016/J.JECP.2017.09.017

Henderson, L. M., Weighall, A. R., Brown, H., & Gareth Gaskell, M. (2012). Consolidation of vocabulary is associated with sleep in children. *Developmental Science*, *15*(5), 674–687. https://doi.org/10.1111/j.1467-7687.2012.01172.x

Hoedlmoser, K., Heib, D. P. J., Roell, J., Peigneux, P., Sadeh, A., Gruber, G., & Schabus, M. (2014). Slow Sleep Spindle Activity, Declarative Memory, and General Cognitive Abilities in Children. *Sleep*, *37*(9), 1501–1512. https://doi.org/10.5665/sleep.4000

Hopkins, R. F., Lyle, K. B., Hieb, J. L., & Ralston, P. A. S. (2016). Spaced Retrieval Practice Increases College Students' Short- and Long-Term Retention of Mathematics Knowledge. *Educational Psychology Review*, *28*(4), 853–873. https://doi.org/10.1007/S10648-015-9349-8/TABLES/4

Horváth, K., Myers, K., Foster, R. and Plunkett, K. (2015), Napping facilitates word learning in early lexical development. *J Sleep Res*, 24: 503-509. https://doi.org/10.1111/jsr.12306

Imundo, M. N., Pan, S. C., Bjork, E. L., & Bjork, R. A. (2021). Where and how to learn: The interactive benefits of contextual variation, restudying, and retrieval practice for learning. *Quarterly Journal of Experimental Psychology*, *74*(3), 413–424. https://doi.org/10.1177/1747021820968483

JASP Team (2022). JASP (Version 0.16.3)[Computer software].

James, E. (2019) *Understanding individual differences in learning and consolidating new vocabulary*. PhD thesis, University of York.

James, E., Gaskell, M. G., & Henderson, L. M. (2019). Offline consolidation supersedes prior knowledge benefits in children's (but not adults') word learning. *Developmental Science*, *22*(3), e12776. https://doi.org/10.1111/desc.12776

James, E., Gaskell, M. G., Weighall, A., & Henderson, L. (2017). Consolidation of vocabulary during sleep: The rich get richer? *Neuroscience & Biobehavioral Reviews*, *77*, 1–13. https://doi.org/10.1016/J.NEUBIOREV.2017.01.054

Jeffreys, H. (1998). *The theory of probability*. OUP Oxford.

Joober, R., Schmitz, N., Annable, L., & Boksa, P. (2012). Publication bias: what are the challenges and can they be overcome?. *Journal of Psychiatry and Neuroscience*, *37*(3), 149-152.

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

Kang, S. H. K., Mcdermott, K. B., & Roediger, H. L. (2007). *Test format and corrective feedback modify the effect of testing on long-term retention*. https://doi.org/10.1080/09541440601056620

Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1250–1257. https://doi.org/10.1037/a0023436

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*(6018), 772–775. https://doi.org/10.1126/SCIENCE.1199327/SUPPL_FILE/KARPICKE_SOM.PDF

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-Based Learning: An Episodic Context Account. *Psychology of Learning and Motivation - Advances in Research and Theory, 61*, 237–284. https://doi.org/10.1016/B978-0-12-800283-4.00007-1

Karpicke, J. D., & Roediger, H. L. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(4), 704–719. https://doi.org/10.1037/0278-7393.33.4.704

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319*(5865), 966–968. https://doi.org/10.1126/science.1152408

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773-795.

Kelter, R. Bayesian alternatives to null hypothesis significance testing in biomedical research: a non-technical introduction to Bayesian inference with JASP. *BMC Med Res Methodol* 20, 142 (2020). https://doi.org/10.1186/s12874-020-00980-6

Kim, A. S. N., Minooei Saberi, F., Wiseheart, M., & Rosenbaum, R. S. (2018). Ameliorating Episodic Memory Deficits in a Young Adult With Developmental (Congenital) Amnesia. *Journal of the International Neuropsychological Society*, *24*(9), 1003–1012. https://doi.org/10.1017/S1355617718000589

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*, 1–13. https://doi.org/10.3758/s13428-018-1184-7

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. https://doi.org/10.1016/J.JARMAC.2020.07.007

Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology*, *23*(9), 1297–1317. https://doi.org/10.1002/acp.1537

Kornell, N., Klein, P. J., & Rawson, K. A. (2015). Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. *Journal of experimental psychology. Learning, memory, and cognition*, *41*(1), 283–294. https://doi.org/10.1037/a0037850

Kornell, N., & Vaughn, K. E. (2016). How Retrieval Attempts Affect Learning: A Review and Synthesis. *Psychology of Learning and Motivation - Advances in Research and Theory*, *65*, 183–215. https://doi.org/10.1016/BS.PLM.2016.03.003

Kornmeier, J., Sosic-Vasic, Z., & Joos, E. (2022). Spacing learning units affects both learning and forgetting. *Trends in Neuroscience and Education*, *26*, 100173. https://doi.org/10.1016/J.TINE.2022.100173

Kroneisen, M., & Kuepper-Tetzel, C. E. (2021). Using day and night – scheduling retrieval practice and sleep. *Psychology Learning and Teaching*, *20*(1), 40–57. https://doi.org/10.1177/1475725720965363

Kruschke, J.K., Liddell, T.M. The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. Psychon Bull Rev 25, 178–206 (2018). https://doi.org/10.3758/s13423-016-1221-4

Kurdziel, L., Duclos, K., & Spencer, R. M. C. (2013). Sleep spindles in midday naps enhance learning in preschool children. *Proceedings of the National Academy of Sciences of the*

*United States of America*, *110*(43), 17267–17272. https://doi.org/10.1073/pnas.1306418110

Kurdziel, L. B. F., & Spencer, R. M. C. (2016). Consolidation of novel word learning in native English-speaking adults. *Memory*, *24*(4), 471–481. https://doi.org/10.1080/09658211.2015.1019889

Kurth, S., Ringli, M., Geiger, A., LeBourgeois, M., Jenni, O. G., & Huber, R. (2010). Mapping of Cortical Activity in the First Two Decades of Life: A High-Density Sleep Electroencephalogram Study. *Journal of Neuroscience*, *30*(40). http://www.jneurosci.org/content/30/40/13211

Küpper-Tetzel, C. E., Kapler, I. V., & Wiseheart, M. (2014). Contracting, equal, and expanding learning schedules: the optimal distribution of learning sessions depends on retention interval. *Memory & cognition*, *42*(5), 729–741. https://doi.org/10.3758/s13421-014-0394-1

Lahl, O., Wispel, C., Willigens, B., & Pietrowsky, R. (2008). An ultra short episode of sleep is sufficient to promote declarative memory performance. *Journal of Sleep Research*, *17*(1), 3–10. https://doi.org/10.1111/J.1365-2869.2008.00622.X

Landauer, T. K. (1969). Reinforcement as consolidation. *Psychological Review*, *76*(1), 82–96. https://doi.org/10.1037/h0026746

Latimier, A., Peyre, H., & Ramus, F. (2021). A meta-analytic review of the benefit of spacing out retrieval practice episodes on retention. *Educational Psychology Review*, *33*(3), 959–987. https://doi.org/10.1007/s10648-020-09572-8

Law, J., Rush, R., Schoon, I., & Parsons, S. (2009). Modeling developmental language difficulties from school entry into adulthood: literacy, mental health, and employment outcomes. *Journal of speech, language, and hearing research: JSLHR*, *52(6*), 1401–1416. https://doi.org/10.1044/1092-4388(2009/08-0142)

Leach, L., & Samuel, A. G. (2007). Lexical Configuration and Lexical Engagement: When Adults Learn New Words. *Cognitive Psychology*, *55*(4), 306. https://doi.org/10.1016/J.COGPSYCH.2007.01.001

Lee, J.-H., Lim, Y., Wiederhold, B. K., & Graham, S. J. (2005). A Functional Magnetic Resonance Imaging (fMRI) Study of Cue-Induced Smoking Craving in Virtual Environments. *Applied Psychophysiology and Biofeedback*, *30*(3), 195–204. https://doi.org/10.1007/s10484-005-6377-z

Leonard, L. B., Deevy, P., Karpicke, J. D., Christ, S. L., & Kueser, J. B. (2020). After Initial Retrieval Practice, More Retrieval Produces Better Retention Than More Study in the Word Learning of Children With Developmental Language Disorder. *Journal of Speech, Language, and Hearing Research*, *63*(8), 2763–2776. https://doi.org/10.1044/2020_JSLHR-20-00105

Lindsay, S., & Gaskell, M. G. (2013). Lexical integration of novel words without sleep. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 608–622. https://doi.org/10.1037/a0029243

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. https://doi.org/10.1080/13825580701322171

Lokhandwala, S., & Spencer, R. M. C. (2021). Slow wave sleep in naps supports episodic memories in early childhood. *Developmental Science*, *24*(2), e13035. https://doi.org/10.1111/DESC.13035

Ly, A., Verhagen, J., & Wagenmakers, E. J. (2016). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, *72*, 19–32. https://doi.org/10.1016/J.JMP.2015.06.004

Macchitella, L., Marinelli, C. V., Signore, F., Ciavolino, E., & Angelelli, P. (2020). Sleepiness, Neuropsychological Skills, and Scholastic Learning in Children. *Brain Sciences, 10*(8), 1–19. https://doi.org/10.3390/BRAINSCI10080529

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. https://doi.org/10.1080/20445911.2016.1181637

Marshall, L., Helgadóttir, H., Mölle, M., & Born, J. (2006). Boosting slow oscillations during sleep potentiates memory. *Nature 2006 444:7119*, *444*(7119), 610–613. https://doi.org/10.1038/nature05278

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: an open-source, graphical experiment builder for the social sciences. Behavior research methods, 44(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

McClelland, J. L. (2013). Incorporating rapid neocortical learning of new schema-consistent information into complementary learning systems theory. *Journal of Experimental Psychology: General*, *142*(4), 1190–1210. https://doi.org/10.1037/A0033812

McClelland, J. L., McNaughton, B. L., & Lampinen, A. K. (2020). Integration of new information in memory: new insights from a complementary learning systems perspective. *Philosophical Transactions of the Royal Society B*, *375*(1799). https://doi.org/10.1098/RSTB.2019.0637

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457. https://doi.org/10.1037/0033-295X.102.3.419

Mcdermott, K.B. Paradoxical effects of testing: Repeated retrieval attempts enhance the likelihood of later accurate and false recall. *Memory & Cognition 34*, 261–267 (2006). https://doi.org/10.3758/BF03193404

Metcalfe, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory & Cognition 2009 37:8*, *37*(8), 1077–1087. https://doi.org/10.3758/MC.37.8.1077

Menghini, D., Carlesimo, G. A., Marotta, L., Finzi, A., & Vicari, S. (2010). Developmental Dyslexia and Explicit Long-Term Memory. *DYSLEXIA*, *16*, 213–225. https://doi.org/10.1002/dys.410

Miles, S. W. (2014). Spaced vs. massed distribution instruction for L2 grammar learning. *System*, *42*(1), 412–428. https://doi.org/10.1016/J.SYSTEM.2014.01.014

Miley, A. Å., Kecklund, G., & Åkerstedt, T. (2016). Comparing two versions of the Karolinska Sleepiness Scale (KSS). *Sleep and biological rhythms*, *14*(3), 257–260. https://doi.org/10.1007/s41105-016-0048-8

Milton, J. & Treffers-Daller, J. (2013). Vocabulary size revisited: the link between vocabulary size and academic achievement. *Applied Linguistics Review*, *4(1),* 151-172. https://doi.org/10.1515/applirev-2013-0007

Mlinarić, A., Horvat, M., & Šupak Smolčić, V. (2017). Dealing with the positive publication bias: Why you should really publish your negative results. *Biochemia medica*, *27*(3), 030201. https://doi.org/10.11613/BM.2017.030201

Moinzadeh, A. R., Talebinezhad, M. R., & Behazin, A. (2008). Exposure Density in Relation to Learning and Retention in EFL. *The International Journal of Humanities*, *15*(2), 71–98. https://eijh.modares.ac.ir/article-27-6922-en.html

Moreira, B. F. T., Pinto, T. S. S., Starling, D. S. V., & Jaeger, A. (2019). Retrieval Practice in Classroom Settings: A Review of Applied Research. *Frontiers in Education*, *4*. https://doi.org/10.3389/FEDUC.2019.00005

Moss, K. K. H. (1996). Word monitoring. *Language and Cognitive Processes*, *11*(6), 689-694.

Mölle, M., Bergmann, T. O., Marshall, L., & Born, J. (2011). Fast and slow spindles during the sleep slow oscillation: disparate coalescence and engagement in memory processing. *Sleep*, *34*(10), 1411–1421. https://doi.org/10.5665/SLEEP.1290

Mölle, M., Marshall, L., Gais, S., & Born, J. (2002). Grouping of Spindle Activity during Slow Oscillations in Human Non-Rapid Eye Movement Sleep. *Journal of Neuroscience*, *22*(24), 10941–10947. https://doi.org/10.1523/JNEUROSCI.22-24-10941.2002

Newbury, C. R., Crowley, R., Rastle, K., & Tamminen, J. (2021). Sleep deprivation and memory: Meta-analytic reviews of studies on sleep deprivation before and after learning.*Psychological Bulletin*, *147*(11), 1215–1240. https://doi.org/10.1037/bul0000348

Newport, E. L. (1990). Maturational Constraints on Language Learning. *Cognitive Science*, *14*(1), 11–28. https://doi.org/10.1207/S15516709COG1401_2

Ngo, H. V., Fell, J., & Staresina, B. (2020). Sleep spindles mediate hippocampal-neocortical coupling during long-duration ripples. *ELife*, *9*, 1–18. https://doi.org/10.7554/ELIFE.57011

Nicolas, A., Petit, D., Rompré, S., & Montplaisir, J. (2001). Sleep spindle characteristics in healthy subjects of different age groups. *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, *112*(3), 521–527. https://doi.org/10.1016/s1388-2457(00)00556-3

Nir, Y., Staba, R. J., Andrillon, T., Vyazovskiy, V. v., Cirelli, C., Fried, I., & Tononi, G. (2011). Regional Slow Waves and Spindles in Human Sleep. *Neuron*, *70*(1), 153–169. https://doi.org/10.1016/J.NEURON.2011.02.043

Norbury, C. F., Griffiths, H., & Nation, K. (2010). Sound before meaning: word learning in autistic disorders. *Neuropsychologia*, *48*(14), 4012–4019. https://doi.org/10.1016/j.neuropsychologia.2010.10.015

Ohayon, M. M., Carskadon, M. A., Guilleminault, C., & Vitiello, M. v. (2004). Meta-analysis of quantitative sleep parameters from childhood to old age in healthy individuals: developing normative sleep values across the human lifespan. *Sleep*, *27*(7), 1255–1273. http://www.ncbi.nlm.nih.gov/pubmed/15586779

O’Reilly, R.C., Bhattacharyya, R., Howard, M.D. and Ketz, N. (2014), Complementary Learning Systems. *Cognitive Science*, *38*: 1229-1248. https://doi.org/10.1111/j.1551-6709.2011.01214.x

O'Reilly, R. C., & Rudy, J. W. (2000). Computational principles of learning in the neocortex and hippocampus. *Hippocampus*, *10*(4), 389–397. https://doi.org/10.1002/1098-1063(2000)10:4<389::AID-HIPO5>3.0.CO;2-P

Paik, J., and Ritter, F. E. (2015). Evaluating a range of learning schedules: hybrid training schedules may be as good as or better than distributed practice for some tasks. *Ergonomics* 59, 276–290. doi: 10.1080/00140139.2015.1067332

Palma, P., & Titone, D. (2021). Something old, something new: A review of the literature on sleep-related lexicalization of novel words in adults. *Psychonomic bulletin & review*, *28*(1), 96–121. https://doi.org/10.3758/s13423-020-01809-5

Pashler, H., Cepeda, N., Lindsey, R. v., Vul, E., & Mozer, M. C. (2009). Predicting the Optimal Spacing of Study: A Multiscale Context Model of Memory. *Advances in Neural Information Processing Systems*, *22*.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? In *Journal of Experimental Psychology: Learning Memory and Cognition* (Vol. 31, Issue 1, pp. 3–8). https://doi.org/10.1037/0278-7393.31.1.3

Pastötter, B., Bäuml, KH.T. Reversing the testing effect by feedback: Behavioral and electrophysiological evidence. *Cogn Affect Behav Neurosci 16*, 473–488 (2016). https://doi.org/10.3758/s13415-016-0407-6

Pavlik, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, *14*(2), 101–117. https://doi.org/10.1037/1076-898X.14.2.101

Pickering, J. S., Henderson, L. M., & Horner, A. J. (2021). Retrieval practice transfer effects for multielement event triplets. *Royal Society Open Science*, *8*(11). https://doi.org/10.1098/RSOS.201456

Perfetti, C. A., Wlotko, E. W., & Hart, L. A. (2005). Word learning and individual differences in word learning reflected in event-related potentials. *Journal of Experimental*

*Psychology: Learning Memory and Cognition*, *31*(6), 1281–1292. https://doi.org/10.1037/0278-7393.31.6.1281

Petersen-Brown, S., Lundberg, A. R., Ray, J. E., Paz, I. N. dela, Riss, C. L., & Panahon, C. J. (2019). Applying spaced practice in the schools to teach math vocabulary. *Psychology in the Schools*, *56*(6), 977–991. https://doi.org/10.1002/PITS.22248

Petzka, M., Charest, I., Balanos, G. M., & Staresina, B. P. (2020). Does sleep-dependent consolidation favour weak memories? *Cortex*. https://doi.org/10.1016/j.cortex.2020.10.005

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*(4), 437–447. https://doi.org/10.1016/j.jml.2009.01.004

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

Rasch, B., & Born, J. (2013). About Sleep's Role in Memory. *Physiological Reviews*, *93*(2). https://doi.org/10.1152/physrev.00032.2012

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General*, *140*(3), 283–302. https://doi.org/10.1037/a0023956

Rea, C. P., & Modigliani, V. (1985). The effect of expanded versus massed practice on the retention of multiplication facts and spelling lists. *Human Learning: Journal of Practical Research & Applications*, *4*(1), 11–18.

Reifman, J., Kumar, K., Khitrov, M. Y., Liu, J., & Ramakrishnan, S. (2018). PC-PVT 2.0: An updated platform for psychomotor vigilance task testing, analysis, prediction, and visualization. *Journal of neuroscience methods*, *304*, 39-45.

Roediger, H. L., Agarwal, P. K., Mcdaniel, M. A., & Mcdermott, K. B. (2011). *Test-Enhanced Learning in the Classroom: Long-Term Improvements From Quizzing*. https://doi.org/10.1037/a0026252

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. In *Trends in Cognitive Sciences* (Vol. 15, Issue 1, pp. 20–27). Elsevier Current Trends. https://doi.org/10.1016/j.tics.2010.09.003

Roediger III, H. L., Jacoby, J. D., & McDermott, K. B. (1996). Misinformation Effects in Recall: Creating False Memories through Repeated Retrieval. *Journal of Memory and Language*, *35*(2), 300–318. https://doi.org/10.1006/JMLA.1996.0017

Roediger, H. L., & Karpicke, J. D. (2006). The Power of Testing Memory: Basic Research and Implications for Educational Practice. *Perspectives on Psychological Science*, *1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Rossi M., Martin-Chang S., & Ouellette G., (2019) Exploring the Space Between Good and Poor Spelling: Orthographic Quality and Reading Speed, *Scientific Studies of Reading*, 23:2, 192-201, DOI: 10.1080/10888438.2018.1508213

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374. https://doi.org/10.1016/J.JMP.2012.08.001

Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Sarasso, S., Proserpio, P., Pigorini, A., Moroni, F., Ferrara, M., de Gennaro, L., de Carli, F., lo Russo, G., Massimini, M., & Nobili, L. (2014). Hippocampal sleep spindles preceding neocortical sleep onset in humans. *NeuroImage*, *86*, 425–432. https://doi.org/10.1016/J.NEUROIMAGE.2013.10.031

Schimke, E. A. E., Angwin, A. J., Cheng, B. B. Y., & Copland, D. A. (2021). The effect of sleep on novel word learning in healthy adults: A systematic review and meta-analysis. *Psychonomic Bulletin and Review*, *28*(6), 1811–1838. https://doi.org/10.3758/S13423-021-01980-3/TABLES/6

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.

Schmidt, C., Peigneux, P., Muto, V., Schenkel, M., Knoblauch, V., Münch, M., Quervain, D. J.-F. de, Wirz-Justice, A., & Cajochen, C. (2006). Encoding Difficulty Promotes Postlearning Changes in Sleep Spindle Activity during Napping. *Journal of Neuroscience*, *26*(35), 8976–8982. https://doi.org/10.1523/JNEUROSCI.2464-06.2006

Schmid, D., & Stanton, N. A. (2020). Exploring Bayesian analyses of a small-sample-size factorial design in human systems integration: the effects of pilot incapacitation. *Human-Intelligent Systems Integration 2020 1:2*, *1*(2), 71–88. https://doi.org/10.1007/S42454-020-00012-0

Schoch, S. F., Cordi, M. J., & Rasch, B. (2017). Modulating influences of memory strength and sensitivity of the retrieval test on the detectability of the sleep consolidation effect. *Neurobiology of learning and memory*, *145*, 181–189. https://doi.org/10.1016/j.nlm.2017.10.009

Schwartz, R., & Terrell, B. (1983). The role of input frequency in lexical acquisition. *Journal of Child Language, 10*(1), 57-64. doi:10.1017/S0305000900005134

Scullin, M. K., Fairley, J., Decker, M. J., & Bliwise, D. L. (2017). The Effects of an Afternoon Nap on Episodic Memory in Young and Older Adults. *Sleep*, *40*(5). https://doi.org/10.1093/SLEEP/ZSX035

Seabrook, R., Brown, G. D. A., & Solity, J. E. (2005). Distributed and massed practice: from laboratory to classroom. *Applied Cognitive Psychology*, *19*(1), 107–122. https://doi.org/10.1002/acp.1066

Shahid, A., Wilkinson, K., Marcu, S., & Shapiro, C. M. (2011). Karolinska sleepiness scale (KSS). In *STOP, THAT and one hundred other sleep scales* (pp. 209-210). Springer, New York, NY.

Shea, C. H., Lai, Q., Black, C., & Park, J.-H. (2000). Spacing practice sessions across days benefits the learning of motor skills. *Human Movement Science, 19*(5), 737–760. https://doi.org/10.1016/S0167-9457(00)00021-X

Simone, P. M., Bell, M. C., & Cepeda, N. J. (2013). Diminished But Not Forgotten: Effects of Aging on Magnitude of Spacing Effect Benefits. *The Journals of Gerontology: Series B, 68*(5), 674–680. https://doi.org/10.1093/GERONB/GBS096

Smith, F. R., Gaskell, M. G., Weighall, A. R., Warmington, M., Reid, A. M., & Henderson, L. M. (2018). Consolidation of vocabulary is associated with sleep in typically developing children, but not in children with dyslexia. *Developmental Science*, *21*(5), 12639. https://doi.org/10.1111/desc.12639

Smith, S. M., Glenberg, A., & Bjork, R. A. (1978). Environmental context and human memory. *Memory & Cognition 1978 6:4*, *6*(4), 342–353. https://doi.org/10.3758/BF03197465

Smith, T. A., & Kimball, D. R. (2010). Learning from feedback: Spacing and the delay–retention effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 80–95. https://doi.org/10.1037/a0017407

Smith, C. D., & Scarf, D. (2017). Spacing Repetitions Over Long Timescales: A Review and a Reconsolidation Explanation. *Frontiers in Psychology*, *8*, 962. https://doi.org/10.3389/fpsyg.2017.00962

Sobel, H. S., Cepeda, N. J., & Kapler, I. v. (2011). Spacing effects in real-world classroom vocabulary learning. *Applied Cognitive Psychology*, *25*(5), 763–767. https://doi.org/10.1002/acp.1747

Son, L.K., Simon, D.A. (2012) Distributed Learning: Data, Metacognition, and Educational Implications. *Educ Psychol Rev 24*, 379–399 (2012). https://doi.org/10.1007/s10648-012-9206-y

Spanò, G., Gómez, R. L., Demara, B. I., Alt, M., Cowen, S. L., & Edgin, J. O. (2018). REM sleep in naps differentially relates to memory consolidation in typical preschoolers and children with Down syndrome. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(46), 11844–11849. https://doi.org/10.1073/pnas.1811488115

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology, 30(9)*, 641–656. https://doi.org/10.1037/h0063404

Staresina, B. P., Bergmann, T. O., Bonnefond, M., van der Meij, R., Jensen, O., Deuker, L., Elger, C. E., Axmacher, N., & Fell, J. (2015). Hierarchical nesting of slow oscillations, spindles and ripples in the human hippocampus during sleep. *Nature Neuroscience*, *18*(11), 1679–1686. https://doi.org/10.1038/nn.4119

Steber, S., & Rossi, S. (2021). The challenge of learning a new language in adulthood: Evidence from a multi-methodological neuroscientific approach. *PLOS ONE*, *16*(2), e0246421. https://doi.org/10.1371/JOURNAL.PONE.0246421

Steriade, M. (2006). Grouping of brain rhythms in corticothalamic systems. *Neuroscience*, *137*(4), 1087–1106. https://doi.org/10.1016/J.NEUROSCIENCE.2005.10.029

Stickgold, R., & Walker, M. P. (2007). Sleep-dependent memory consolidation and reconsolidation. *Sleep Medicine*, *8*(4), 331–343. https://doi.org/10.1016/J.SLEEP.2007.03.011

Surma, T., Vanhoyweghen, K., Camp, G., & Kirschner, P. A. (2018). The coverage of distributed practice and retrieval practice in Flemish and Dutch teacher education textbooks. *Teaching and Teacher Education, 74,* 229-237. https://doi.org/10.1016/j.tate.2018.05.007

Sutterer, D. W., & Awh, E. (2016). Retrieval practice enhances the accessibility but not the quality of memory. Psychonomic bulletin & review, 23(3), 831–841. https://doi.org/10.3758/s13423-015-0937-x

Suzuki, Y. and DeKeyser, R. (2015), Comparing Elicited Imitation and Word Monitoring as Measures of Implicit Knowledge. Language Learning, 65: 860-895. https://doi.org/10.1111/lang.12138

Swinson, J. (2013). British Ability Scales 3.

Takashima, A., Bakker, I., van Hell, J. G., Janzen, G., & McQueen, J. M. (2017). Interaction between episodic and semantic memory networks in the acquisition and consolidation of novel spoken words. *Brain and language*, *167*, 44–60. https://doi.org/10.1016/j.bandl.2016.05.009

Tamminen, J., Lambon Ralph, M. A., & Lewis, P. A. (2013). The Role of Sleep Spindles and Slow-Wave Activity in Integrating New Information in Semantic Memory. *Journal of Neuroscience*, *33*(39). http://www.jneurosci.org/content/33/39/15376

Tamminen, J., Payne, J. D., Stickgold, R., Wamsley, E. J., & Gaskell, M. G. (2010). Sleep Spindle Activity is Associated with the Integration of New Memories and Existing Knowledge. *Journal of Neuroscience*, *30*(43). http://www.jneurosci.org/content/30/43/14356

Timofeev, I., Grenier, F., Bazhenov, M., Sejnowski, T. J., & Steriade, M. (2000). Origin of Slow Cortical Oscillations in Deafferented Cortical Slabs. *Cerebral Cortex*, *10*(12), 1185–1199. https://doi.org/10.1093/CERCOR/10.12.1185

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: Questions and answers. *Experimental Psychology*, *56*(4), 252–257. https://doi.org/10.1027/1618-3169.56.4.252

Tse, D., Langston, R. F., Kakeyama, M., Bethus, I., Spooner, P. A., Wood, E. R., Witter, M. P., & Morris, R. G. M. (2007). Schemas and memory consolidation. *Science*, *316*(5821), 76–82. https://doi.org/10.1126/SCIENCE.1135935

Tse, D., Takeuchi, T., Kakeyama, M., Kajii, Y., Okuno, H., Tohyama, C., Bito, H., & Morris, R. G. M. (2011). Schema-dependent gene activation and memory encoding in neocortex. *Science*, *333*(6044), 891–895. https://doi.org/10.1126/SCIENCE.1205274

Tucker, M. A., & Fishbein, W. (2008). Enhancement of Declarative Memory Performance Following a Daytime Nap Is Contingent on Strength of Initial Task Acquisition. *Sleep*, *31*(2), 197–203. https://doi.org/10.1093/SLEEP/31.2.197

van den Bergh, D., Van Doorn, J., Marsman, M., Draws, T., Van Kesteren, E. J., Derks, K., ... & Wagenmakers, E. J. (2020). A tutorial on conducting and interpreting a Bayesian ANOVA in JASP. *LAnnee psychologique*, *120*(1), 73-96.

Van Der Werf, Y. D., Altena, E., Schoonheim, M. M., Sanz-Arigita, E. J., Vis, J. C., De Rijke, W., & Van Someren, E. J. (2009). Sleep benefits subsequent hippocampal functioning. *Nature neuroscience*, *12*(2), 122–123. https://doi.org/10.1038/nn.2253

van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., Etz, A., Evans, N. J., Gronau, Q. F., Haaf, J. M., Hinne, M., Kucharský, Š., Ly, A., Marsman, M., Matzke, D., Gupta, A. R. K. N., Sarafoglou, A., Stefan, A., Voelkel, J. G., & Wagenmakers, E.-J. (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*, *28*(3), 813–826. https://doi.org/10.3758/s13423-020-01798-5

van Rijn, E., Gouws, A., Walker, S., Knowland, V. C. P., Cairney, S. A., Gaskell, M. G., & Henderson, L. M. (2021). Do naps benefit novel word learning? Developmental differences and white matter correlates. *BioRxiv*, 2021.11.22.469237. https://doi.org/10.1101/2021.11.22.469237

Vilberg, K. L., & Davachi, L. (2013). Perirhinal-hippocampal connectivity during reactivation is a marker for object-based memory consolidation. *Neuron*, *79*(6), 1232–1242. https://doi.org/10.1016/j.neuron.2013.07.013

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. https://doi.org/10.1016/J.JECP.2019.02.003

Vlach, H. A., & Johnson, S. P. (2013). Memory constraints on infants' cross-situational statistical learning. *Cognition*, *127*(3), 375–382. https://doi.org/10.1016/J.COGNITION.2013.02.015

Vlach, H. A., & Sandhofer, C. M. (2012). Distributing learning over time: the spacing effect in children's acquisition and generalization of science concepts. *Child Development*, *83*(4), 1137–1144. https://doi.org/10.1111/j.1467-8624.2012.01781.x

Vojdanoska, M., Cranney, J. and Newell, B.R. (2010), The testing effect: The role of feedback and collaboration in a tertiary classroom setting. *Appl. Cognit. Psychol.*, *24*: 1183-1195. https://doi.org/10.1002/acp.1630

Walker, S., Henderson, L. M., Fletcher, F. E., Knowland, V. C. P., Cairney, S. A., & Gaskell, M. G. (2019). Learning to live with interfering neighbours: the influence of time of learning and level of encoding on word learning. *Royal Society Open Science*, *6*(4), 181842. https://doi.org/10.1098/rsos.181842

Wechsler, D. (1945). *Wechsler memory scale.* Psychological Corporation.

Wegener, S., Wang, H. C., Beyersmann, E., Nation, K., Colenbrander, D., & Castles, A. (2022). The effects of spacing and massing on children's orthographic learning. *Journal of Experimental Child Psychology*, *214*, 105309. https://doi.org/10.1016/J.JECP.2021.105309

Wegener, S., Wang, H. C., Beyersmann, E., Nation, K., Colenbrander, D., & Castles, A. (2022). The effects of spacing and massing on children's orthographic learning. *Journal of*

*Experimental Child Psychology, 214,* 105309. https://doi.org/10.1016/J.JECP.2021.105309

Weighall, A. R., Henderson, L. M., Barr, D. J., Cairney, S. A., & Gaskell, M. G. (2017). Eye-tracking the time-course of novel word learning and lexical competition in adults and children. *Brain and Language*, *167*, 13–27. https://doi.org/10.1016/j.bandl.2016.07.010

Westfall, P. H., Johnson, W. O., & Utts, J. M. (1997). A Bayesian Perspective on the Bonferroni Adjustment. *Biometrika*, *84*(2), 419–427. http://www.jstor.org/stable/2337467

Wilhelm, I., Rose, M., Imhof, K. I., Rasch, B., Büchel, C., & Born, J. (2013). The sleeping child outplays the adult's capacity to convert implicit into explicit knowledge. *Nature Neuroscience*, *16*(4), 391–393. https://doi.org/10.1038/nn.3343

Wilhelm, I., Diekelmann, S., & Born, J. (2008). Sleep in children improves memory performance on declarative but not procedural tasks. *Learning & Memory (Cold Spring Harbor, N.Y.)*, *15*(5), 373–377. https://doi.org/10.1101/lm.803708

Williams, S. E., & Horst, J. S. (2014). Goodnight book: Sleep consolidation improves word learning via storybooks. *Frontiers in psychology, 5,* 184. https://doi.org/10.3389/FPSYG.2014.00184

Wilson, M. A., & McNaughton, B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science*, *265*(5172), 676–679. https://doi.org/10.1126/SCIENCE.8036517

He, A. X., Huang, S., Waxman, S., & Arunachalam, S. (2020). Two-year-olds consolidate verb meanings during a nap. *Cognition*, *198*, 104205. https://doi.org/10.1016/J.COGNITION.2020.104205

Yassa, M. A., & Stark, C. E. L. (2011). Pattern separation in the hippocampus. *Trends in Neurosciences*, *34*(10), 515–525. https://doi.org/10.1016/J.TINS.2011.06.006

Zhuang, L., Wang, J., Xiong, B., Bian, C., Hao, L., Bayley, P. J., & Qin, S. (2021). Rapid neural reorganization during retrieval practice predicts subsequent long-term retention and false memory. *Nature Human Behaviour 2021 6:1*, *6*(1), 134–145. https://doi.org/10.1038/s41562-021-01188-4

Zigterman, J. R., Simone, P. M., & Bell, M. C. (2015). Within-session spacing improves delayed recall in children. *Memory*, *23*(4), 625–632. https://doi.org/10.1080/09658211.2014.915975

Åkerstedt, T., & Gillberg, M. (1990). Subjective and objective sleepiness in the active individual. *The International journal of neuroscience*, *52*(1-2), 29–37. https://doi.org/10.3109/00207459008994241