Spacing and Task Complexity in Mathematics Learning

Ewan Murray

Doctor of Philosophy

University of York

Psychology

November 2025

Abstract

This thesis has two goals, to consider the evidence for the spacing and testing effects in mathematics, and to investigate the effect of task complexity on the spacing effect. Chapter 2 presents a meta-analysis that aimed to synthesize the available evidence for the spacing and testing effects in mathematics learning. I found evidence for a robust effect of spaced practice and that this effect is larger when material is learnt in isolation rather than when embedded in a course, however, there is not yet enough evidence to show a robust testing effect in mathematics material. The experimental chapters of this thesis aimed to investigate the relationship between spacing and task complexity in mathematics learning. As task complexity can vary depending on the individual's prior learning and experiences, this poses a problem when attempting to prescribe an individual with a personalised spaced learning schedule. I aimed to reduce the effect of prior knowledge by creating artificial tasks where each part is known to participants (i.e., the action "divide by two" as a step in a procedure, or the evaluation "is it a multiple of five?" to categorise a number), however, how the parts are structured is manipulated. Chapter 3 presents two experiments where procedural complexity, number of steps in the procedure, was manipulated. I found a consistent main effect of spacing, however spacing did not interact with procedural complexity. Chapter 4 presents two experiments where element interactivity, the number of elements that the learner must hold in working memory at once, was manipulated. This was done by varying the structure of categories. Experiment three found no main effect of spacing, however, I suggest this was due to interference between the categories. Overall, I found no evidence that procedural complexity or element interactivity interacts with the spacing effect and suggest this may be due to competing underlying mechanisms of the spacing effect that enable the effect's robustness across tasks.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to firstly thank my supervisors Silke and Aidan, whose constant support has made the whole PhD process very enjoyable. I enjoyed our weekly meetings tremendously and will really miss our discussions of experiments, theories, and life more generally. I could not have asked for better supervisors over the last three and a bit years.

The Psychology department at the University of York is an exceptional place to undertake a PhD and the academic and support staff created a great environment to research. Some key discussions with Philip Quinlan (whose early advice helped shape the procedures task) and Layla Unger (who introduced me to the concept of relational categories) provided turning points in the design of my experiments. My Thesis Advisory Panel meetings with Beth Jefferies, Daniel Baker and Tom Hartley helped shape all of my work and I really valued our discussions. All the members of the SLAM lab always provided great feedback and interesting new ideas whenever I had new results to share.

I feel very fortunate to have made so many great friends in the department and feel I could very easily just copy and paste in the entire directory from the department website to thank. In particular my office mates: Aida, Adam, and Sophie made a huge impact on the high quality of life I attained during my time here (even when Adam was loudly grinding his coffee beans). Also, the postdocs in the nearby office Jamie, Arianna, and Chloe helped me out a lot and became great friends. The other frequent lunch attendees Cameron, Vanessa, Bee, Dan, Charlotte, Lyndon, were always great to have a chat with and often a quiz. In the wider cohort, I was glad to start at the same time as some great people, Aaron, Hannah, Chen, Ramya, Alex, Lauren and Emily. I was very glad when Ellie joined the department and always loved climbing

and quizzing with her and Zach. Outside of the dept. Alex and Robs were excellent roommates and the climbing society was another excellent source of friends.

I would like to thank my family, in particular my parents, who have continued to make the trip up to York on a regular basis and treat their son, the perpetual student, to some good food and company.

Finally, I have to thank Lilly, who I had the great luck to meet pretty much on the first day of my PhD and who provided endless support throughout the whole process. It would have been much less fun without you to come home to.

I first came to the University of York in 2016 and, except for a brief foray into teaching, I was very glad to consider the campus a second home. I will miss it dearly.

Author's Declaration

I declare that this thesis is a presentation of original work, and I am the sole author. This work has not previously been presented for a degree or other qualification at this University or elsewhere. All sources are acknowledged as references. As this is a ***journal style thesis***, the beginnings of chapter 2, 3 and 4 have an additional section outlining the publication status of each chapter.

Ewan Murray was supported by a departmental studentship from the University of York.

1 Literature Review

Many recent suggested best practices in education have focused on the long-term retention of key mathematical knowledge (Ofsted, 2021). One reason for this is that the majority of mathematics learning builds directly on previous ideas, which may have been introduced days, months, or years earlier. Unfortunately, students often struggle to retrieve this knowledge (Karpicke, 2012). Learners' inability to retrieve past material requires instructors to allocate additional time reintroducing it, which impedes not only the learners' own progress but also, in classroom settings, the progress of all. As time is a scarce resource in educational settings, making this an important problem to remedy.

Fortunately, there are promising interventions that require little additional time and resources: distributing practice and promoting active retrieval. These two interventions employ robust phenomena from cognitive psychology: the spacing effect and testing effect. The spacing effect describes the difference in retention of information when practice is distributed over multiple sessions, rather than in one massed session. The testing effect describes the change in retention when material is actively retrieved rather than restudied. In combination, they form spaced retrieval practice (Hopkins et al., 2016). Already, spaced retrieval practice is in regular use by many students in the United Kingdom. Online Platforms such as ARC Education (n.d.) and Sparx Maths (n.d.) claim to harness the spacing and testing effect to improve retention for pupils. For example, *Sparx Maths (n.d.)* states that they "*ensure the practice uses spaced repetition and interleaving to support a change in students' long-term memories*." This is in reference to a technical report commissioned by Sparx and implemented by RAND Europe and the University of Cambridge (Brown et al., 2021). They found that greater time spent on the Sparx Maths platform is positively and significantly associated with higher outcomes in maths.

However, there is still limited research into the efficacy of these programs with no peer-reviewed studies investigating their use of spacing (Gidaropoulos, 2021). With sites such as Sparx Maths being in use by more than two million students in more than 2,200 schools (Sparx Maths, n.d.), a small positive effect on retention could have a large impact. Furthermore, there is an increasing push to bring these effects into everyday classroom use. A recent report by the British government suggests frequent, low stakes testing as an essential strategy to improve outcomes (Ofsted, 2021). Therefore, understanding when, how and what factors moderate the use of spaced retrieval, before further widespread use, is essential.

This review has two purposes. Firstly, to review the existing evidence for spaced retrieval practice for mathematics and, secondly, to explore factors that moderate its effectiveness. The evidence for the spacing and testing effect will be discussed, and different theoretical accounts will be described and evaluated, followed by a potential problem with investigating retrieval in mathematics. Next, the evidence on the effectiveness of combining spacing and testing (spaced retrieval practice) in mathematics will be reviewed along with proposed moderators. Subsequently, the discussion will turn to the concept of complexity, its definition, importance, and potential influence on the efficacy of spaced retrieval practice, as well as how it may interact with different theoretical accounts. Finally, I explain how the current literature motivated my hypotheses.

1.1 Spacing Effect and Distributed Practice

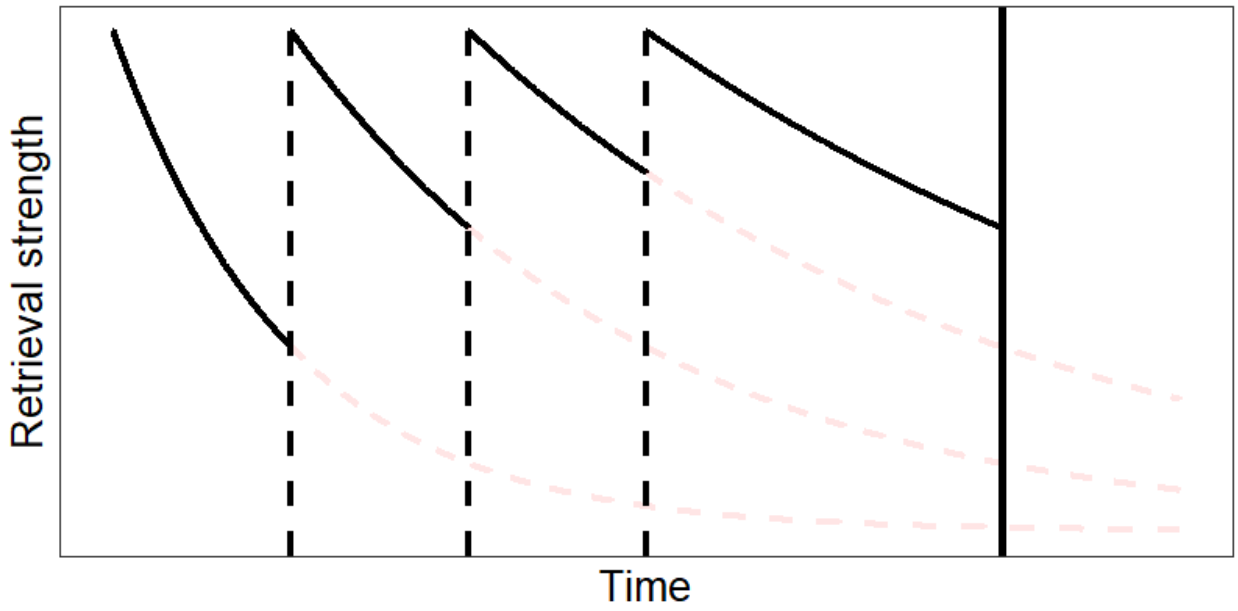
Multiple phrases are used in the literature to describe the difference in retention (i.e., performance on a delayed post-test) when practice is spaced over multiple sessions. The spacing effect refers to the observed change in retention when learning is spaced over multiple sessions, rather than massed into a single session (Delaney et al., 2010). The lag effect is the change in

retention when two different spacing routines are compared (Küpper-Tetzel, 2014). Distributed practice is often used as an umbrella term for interventions that utilise the spacing or lag effects (Benjamin & Tullis, 2010). A meta-meta-analysis by Hattie (2008) into different types of educational interventions found spaced versus massed practice to be one of the most effective. They found a mean effect size of $d = 0.71$, using two meta-analyses containing 63 studies and 5,028 participants. To illustrate what $d = 0.7$ means, consider a study aimed at increasing IQ (where the mean IQ is 100 and standard deviation is 15). An effect size of 0.7 would correspond to a 10.5-point increase in IQ.

The study of the spacing effect began in parallel with the advent of experimental psychology. Ebbinghaus (1913) memorised nonsense syllables and plotted forgetting curves, finding that memories decay exponentially (i.e., quickly at first and then slower). Importantly, this rate of decay slows down after subsequent recall/ retrieval (see Figure 1-1); these results have been replicated successfully (Murre & Dros, 2015). The spacing effect has been the subject of hundreds of experiments (Cepeda et al., 2006; Hattie, 2008), proving to be a robust effect found across age groups, tasks, and species (Walsh et al., 2018). However, most prior experiments on the spacing effect have investigated learning verbal material (Delaney et al., 2010, provides a critical review).

Figure 1-1

Forgetting Curve Showing the Change in Retrieval Strength as a Function of Time



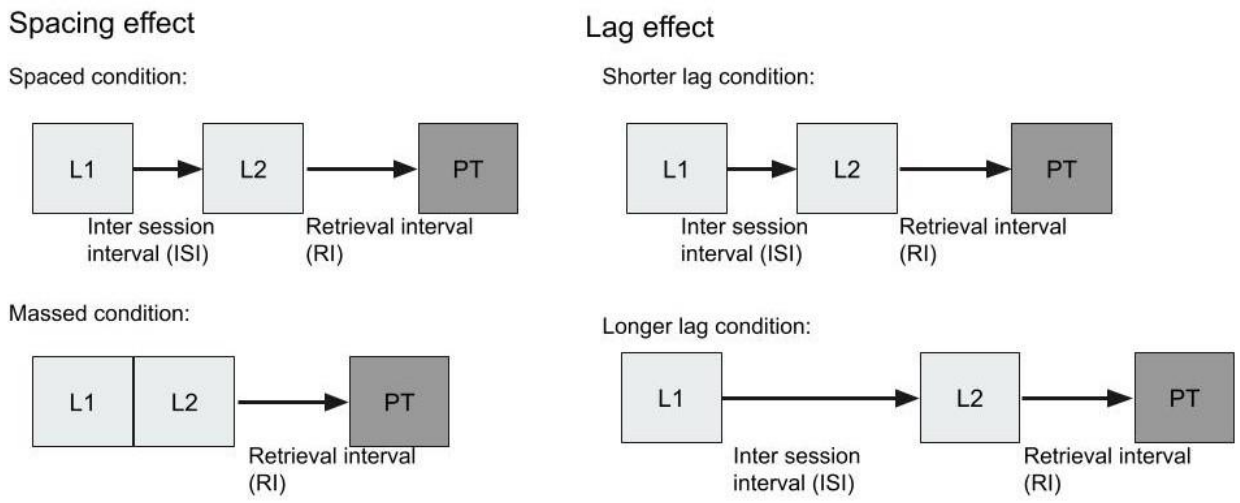
Note. Retrieval strength signifies how easily the item can be retrieved from long-term memory. The thick vertical line represents a post-test. The dashed vertical lines signify retrieval events, the lighter dashed lines show how the memory would have continued to decay without the retrieval events.

Spacing effect experiments can be conducted either within a single session or across multiple sessions (Küpper-Tetzel, 2014). Within-session experiments look at the temporal spacing of stimuli within a single session, whilst across-session experiments look at the spacing of practice across multiple sessions. Within mathematics, there is little work done on within-session spacing (exceptions being: Foster et al., 2019; Rea & Modigliani, 1985), while this is much more common within other domains such as verbal learning (Delaney et al., 2010). Therefore, this review will focus on across-session spacing experiments. In the most basic

across-session spacing effect experiment (see Figure 1-2), retention on a post-test is compared against two practice conditions: practice is either massed into one session or distributed over two sessions. This involves two key design decisions: the inter-session interval (i.e., the time between the initial and subsequent learning sessions) and the retrieval interval (i.e., the time from the final learning session to the post-test). The amount and type of practice are kept constant, therefore, subject to adequate randomisation, the only difference between the two conditions is the temporal spacing.

Figure 1-2

Spacing and Lag Effect Experiments



Note. Diagram showing basic spacing effect and lag effect paradigms. The spacing effect is a special case of the lag effect where the inter-session interval is zero. A simple spacing experiment design consisting of a massed condition where all practice is undertaken in one session and a spaced condition where practice is split into two sessions with an inter-session interval between the first and second learning events. Both conditions are followed by a delayed test.

The optimum inter-session interval is linked to the length of the required retrieval interval. In one large-scale online study (N = 1,350), they taught participants 32 obscure facts and varied the inter-session interval from 7 to 105 days and the retrieval interval from 7 to 350 days (Cepeda et al., 2008). It was found that the optimum inter-session interval depended on the retrieval interval required. For example, to be able to recall a fact 35 days later they found it was best to wait 8 days after initial learning to retrieve it, but to recall a fact 350 days later the optimum inter-session interval was 27 days. This relationship between the inter-session and retrieval interval has been labelled the Glenberg surface (Delaney et al., 2010). It refers specifically to the non-monotonic relationship between the length of the inter-session and retrieval interval: increasing the inter-session interval increases retention up to a point after which retention begins to fall again. This means there is no one "optimum" spacing schedule, but rather it depends on how long the learner is required to remember an item.

More complex spacing schedules are possible with more than two sessions. In this case, key design features are the number of sessions and whether these sessions are uniformly spaced or expanding. This was an interesting question to ask as initial short gaps boost the chance of a successful retrieval, which strengthens the memory allowing for a greater chance of retrieval after the next longer gap (Rea & Modigliani, 1985). Secondly, increasing the gaps should increase the difficulty of retrieval, and more effort should result in greater gains in retrieval strength (Bjork et al., 2011). However, a recent meta-analysis looking at spaced retrieval practice found no significant difference between expanding and uniform designs (Latimier et al., 2021). On the other hand, if both schedules offer equal benefits, expanding schedules are more time-efficient because they produce the same retention gain over the same period with fewer practice sessions.

Alongside increased retention, spacing also improves students' and teachers' ability to accurately gauge learning. For instance, when Year 7 students were asked to predict their scores on a post-test after completing either a massed or spaced practice routine, those who engaged in spaced practice not only achieved higher scores but also made more accurate predictions (Emeny et al., 2021). In contrast, students who followed the massed schedule tended to be overconfident with their predictions. Emeny et al. (2021) suggest this overconfidence may have arisen because massed practice led to greater fluency within the session; however, this performance did not lead to greater long-term learning.

1.2 Theories of the Spacing Effect

Despite the large body of evidence in support of the spacing effect, there is a lack of consensus on the underlying mechanisms (Delaney et al., 2010; Dempster, 1988; Walsh et al., 2018). Many theories that aim to explain the spacing effect; however, currently no one theory, or combination, adequately explains the phenomena (see Delaney et al. (2010) for a critical review of within-session verbal learning spacing experiments and Küpper-Tetzel (2014) for across-session). The most frequently discussed theories include study-phase retrieval, contextual variability, and deficient processing.

Study-phase retrieval suggests that the memory of a previously studied item is strengthened by the retrieval of the original learning event (Thios & D'Agostino, 1976). The degree to which study-phase retrieval improves retention is dependent on the difficulty of recall (Küpper-Tetzel, 2014). The more difficult the item is to recall - while still being successfully retrieved - the better for future recall. This aligns with the concept of desirable difficulties, which suggests that certain techniques (e.g., spacing, interleaving, retrieval practice) may initially hinder performance but ultimately lead to greater long-term retention (Bjork et al., 2011).

Evidence for this phenomenon was found by Magliero (1983) when spacing caused increased processing effort (measured by pupil dilation) for word pair learning. When the retrieval interval is too long, the probability of successful retrieval decreases, resulting in poorer retention for that item according to the study-phase retrieval account This potentially explains the non-monotonic relationship between inter-session intervals and retrieval intervals. Additional evidence for the study-phase retrieval comes from an experiment showing that participants could judge the spacing between two presentations of the same word but struggled to do so with two differing words (Hintzman et al., 1973). Further support was found by Wahlheim et al. (2014), who asked participants to study two word lists. They found that when participants were asked to indicate whether a word was repeated on either the current or the previous list, their future recall ability of these repetitions was enhanced when the word appeared on the previous list compared to when it was repeated within the same list.

Another account of the spacing effect is *contextual variability* (Glenberg, 1979). This theory suggests that during the initial (and any subsequent) retrieval, contextual information is automatically encoded alongside the learning material and that this information provides additional access routes to aid retrieval. This additional information may be related to the environment the learning took place in, such as the location or smells while learning or even the learner's current state of mind (Küpper-Tetzel, 2014). Küpper-Tetzel (2014) provides several examples of studies that manipulated the variability of the context between the initial and final retrieval and found they showed no significant increase or even led to a decrease in performance (Dempster, 1988). An early, though notable, failure to empirically support contextual variability was performed by Ross and Landauer (1978). In this experiment, participants learnt two lists. One list where an item at position x in a sequence is repeated at position y. A second list where

two different items are positioned at x and y. The probability of recalling one of the repeated words, during free recall, would be the same as the probability of recalling one of the two different words. This is because they are at the same position in the list, therefore have the same context. They did not find this to be the case, therefore considered this evidence against the contextual variability theory. However, Lohnas et al. (2011) ran the same analyses on six different previous free recall experimental data sets and found the relationship described above. This provides some empirical evidence for contextual variability. They go on to link contextual variation with the study-phase retrieval hypothesis as it would make sense that during study-phase retrieval the original contextual information and the contextual information from subsequent repetitions are all encoded providing additional retrieval routes.

The *d*eficient process*ing account* of the spacing effect suggests that the phenomenon arises due to learners not processing the material in sufficient depth under the massed condition (Hintzman, 1974). Early studies found that during self-paced massed practice participants spent less time on material that was previously presented, while they spent longer on spaced items (Shaughnessy et al., 1972). This additional exposure was found in the distributed practice conditions but did not fully account for the gains in the distributed practice condition. More recently, eye tracking studies have lent support to this theory, finding that when items were distributed, they received more attentional processing than the massed items (Koval, 2019). Furthermore, this attention was a significant mediator of the efficacy of spacing. This theory differs from the others as it focuses on the disadvantages of massed practice more than the advantages of spacing and therefore some have argued it is not a true spacing effect theory (Delaney et al., 2010). Other researchers do consider it a potential mechanism for the spacing effect but point out additional weaknesses such as its inability to explain why increasing the gap

between spacing sessions produces a greater effect (Benjamin & Tullis, 2010). Overall, there is evidence that deficient processing may affect the efficacy of distributed practice, however, there is little evidence to support a claim that deficient processing is the sole mechanism for the spacing effect.

Recently, as an offshoot of cognitive load theory literature (Sweller, 1988), *Working Memory Resource Depletion* has been offered as an alternative theory to explain some forms of the spacing effect Chen & Kalyuga (2019). They suggest that during massed practice working memory resources are depleted which leads to reduced performance. While in the spaced condition, participants can rest and restore working memory resources. Importantly, they believe this only works when the material is high in *element interactivity*. In cognitive load theory literature, element interactivity is measured by estimating the number of items required to be held in working memory simultaneously to complete a task (Chen et al., 2023). The important distinction made by this concept is the need for all elements to be held in working memory at once, rather than just requiring more elements to be retrieved i.e., remembering more steps overall. When material is low in element interactivity, they do not find any working memory depletion. Across four experiments, Chen et al. (2024), investigate how element interactivity affects working memory resource depletion and the spacing effect. Their first experiment establishes that material high in element interactivity (a connected passage of text) depletes working memory resources, while low element interactivity tasks do not (disconnected sentences). Next, they show that for low element interactivity material you can find a spacing effect without any working memory resource depletion. These first two experiments suggest that higher element interactivity material depletes working memory resources, which quickly recover over a short rest, while lower element interactivity material does not. However, this effect cannot

account for spacing effects found for low element interactivity material, which is most of the previous findings (i.e., word pairs and lists) (Delaney et al., 2010). They suggest that for low element interactivity material spacing allows for additional rehearsal during rest, which causes the spacing effect.

In Chen et al. (2024)'s third and fourth experiments participants they change the complexity of the material not by altering the material itself, but by manipulating the prior knowledge of the participants. In the third experiment the participants were adults who have never studied calculus before, while in the fourth experiment they were sixth form students who had a solid foundation in calculus but had not learnt the specific rule used in the experiment (the product rule). They hypothesized that the material would be high element interactivity for the novices and low element interactivity for the more experienced learners. They found a significant spacing and working memory resource depletion effect for the novices, but neither effect for the more experienced learners. These results highlight the important link between prior knowledge, element interactivity, and the spacing effect. However, like the deficient processing account a key weakness is their inability to explain why increasing inter-session intervals can boost the efficacy of spacing up to a certain point before levelling off. Similarly, the rest periods in Chen et al. (2024) are short (five minutes) if that is sufficient to restore working memory resources then it cannot explain the relationship between inter-session intervals and performance on the post-test. While working memory resource depletion is a possible mechanism of the spacing effect, it cannot be the sole mechanism.

Numerous models based on the above theories have been created. One experiment compared three computational models, each trained on fourteen previous spacing experiment data sets with a combined sample of 2979 participants (Walsh et al., 2018). The first model,

introduced by the authors, is the Predictive Performance Equation (PPE), which focuses on exponential decay of memories, but adds in a function to allow prior repetitions of the item to reduce the decay rate, which is in line with the study-phase retrieval hypothesis (Hintzman et al., 1973). The second model, first introduced by Pavlik and Anderson (2005), was an extension of ACT-R cognitive architecture, where repetition of a chunk in memory increases its activation, thus making it easier to retrieve. This activation level then decays over time, to represent forgetting. This model is mechanically similar to the PPE and can be considered another formal model of study-phase retrieval. However, the final model, the Search of Associative Memory (Raaijmakers, 2003), combines deficient processing and contextual variability to create a formal model of the spacing effect. The Predictive Performance Equation and Pavlik and Anderson's model performed similarly in predicting the results, while the Search of Associative Memory model performed worse. This provides some evidence for study-phase retrieval and against deficient processing and contextual variability. Overall, there is greatest evidence for study-phase retrieval, possibly in combination with contextual variability.

1.3 Testing Effect and Retrieval Practice

Across the experiments presented in this thesis, I do not manipulate testing versus restudy, however as it is included in the meta-analysis it is briefly discussed here. Retrieval practice is an intervention which harnesses the testing effect. It describes the increase in retention when a to be learned item is actively retrieved from memory as opposed to when it is restudied. Typical retrieval practice paradigms involve an initial learning session followed by a post-test. This initial learning session will either be a retrieval or non-retrieval-based learning task. In a classic retrieval practice study participants read science facts initially and then either reread the facts or practised retrieval through free recall (Roediger & Karpicke, 2006). The rereading group

performed significantly better on a 5-minute delayed test; however, the results reversed on a 1-week post-test. One large meta-analysis containing 159 studies found a medium weighted mean effect size for the testing effect of $g = 0.50$ (CI [0.42, 0.58]), similarly recent meta-analysis looking at the testing effect in the classroom found an effect size ($g = 0.499$)(Yang et al., 2021). To illustrate what a $g = 0.5$ effect means, if a study which claimed to raise IQ (where the mean is 100 and standard deviation is 15) and the control and intervention groups both have equal variances, then a 0.5 effect size would mean a 7.5-point increase in IQ.

1.3.1 Theories of the Testing Effect

Many theories aim to explain the testing effect (Rowland, 2014), but this section will focus on the *retrieval effort hypothesis* (coupled with dual memory theory), the *elaborative retrieval hypothesis*, *transfer-appropriate processing,* and the *episodic context account*. Firstly, the theory sometimes called the retrieval effort hypothesis (also used within the explanations for the spacing effect) suggests that the additional effort required by a more difficult, yet successful, retrieval leads to the testing effect (Pyc & Rawson, 2009). The mechanism in which increased effort leads to increased retrieval is expanded upon by the dual memory theory (Rickard & Pan, 2018). Mechanically, it is almost identical to study-phase retrieval (see section 1.1). They propose that initial learning is encoded in the study phase and subsequent retrieval (with feedback) strengthens the original memory and encodes the test memory. This increases future chances of retrieval in the test condition as the initial or test memory can be retrieved, while in the study condition, only the initial memory can be retrieved. They suggest the additional effort required from retrieval is due to having to encode a new test memory alongside retrieval of the study memory. They formally modelled their theory and evaluated it against the data sets from experiments in their lab and they extracted 114 testing effects from the literature. The model

predicts an envelope, an upper and lower bound, they suspect will contain the magnitude of the testing effect. This envelope captures a third of the range of logical potential testing effects. Of the 144 testing effects, their model's bounds overlapped with the true magnitude of the effect (within the confidence interval) in all but five cases. This provides quantitative evidence for this theory, though their prediction envelope is quite large.

Secondly, the elaborative retrieval hypothesis suggests that the benefit of testing comes from the activation of the target alongside other memories creating an elaborative semantic network increasing the number of pathways that future retrievals can access (Carpenter, 2009). In one experiment, participants learnt word pairs. These pairings were either strongly associated (Toast - Bread) or weakly associated (Basket - Bread) (Carpenter, 2009). They hypothesised that if elaborative retrieval was the mechanism which underlies the testing effect, then items with weak associations will be harder to recall initially, but they will create stronger elaborative routes making final recall better than items with strong associations. They found significant results to support this hypothesis.

Thirdly, transfer-appropriate processing, suggests that retention on the final test benefits from the amount of overlap a retrieval event has with the final test (Blaxton, 1989; Morris et al., 1977). However, an experiment designed to evaluate transfer-appropriate processing against the elaborative retrieval hypothesis found little evidence for transfer-appropriate processing (Carpenter & DeLosh, 2006). In this experiment participants were asked to recall a list of words and the type of retrieval in the training and final test conditions were either the same or were mismatched (i.e., free recall during training then free recall in the final test or multiple choice then free recall). This suggests that similarity of the type of test during retrieval does not reflect

the retention on a post-test of the same or different type, while transfer-appropriate processing would predict that to be the case.

Finally, the episodic context account. Karpicke et al. (2014) outline four assumptions used to create the episodic context account: Firstly, people encode information about the item's temporal context during encoding. Secondly, using any potential cues participants reconstruct the past episodic memory. Thirdly, each retrieval updates the prior representation in long-term memory. Finally, this updated representation allows the subject to limit their search of cues to only the most useful ones. The benefit of testing comes from the updated representation limiting the search to only the features that will help future recall. Support for the episodic context account can be found when participants are asked to make judgements about when they initially studied an item (Whiffen & Karpicke, 2017). In the experiment participants learnt a list of words, then either restudied them or were asked to make a judgement about when they studied the word. The authors hypothesised that making the judgement would cause participants to retrieve the original context of when they studied the word. On a subsequent a free recall task, participants who made the temporal judgements on the words performed significantly better. This supports the episodic context account.

1.4 Spaced Retrieval Practice

Rather than discussing mathematics learning within the spacing or testing literature in isolation, it may be better to consider them both here as it is often difficult to be sure whether students were required to retrieve the information, or not, during practice. Furthermore, there is evidence that spaced retrieval can greatly increase learning. One recent meta-analysis found a large overall mean effect size (g = 1.01, 95% CI [0.68, 1.34]), which when corrected for bias within the literature was still notable (g = 0.74, 95% CI [0.55, 0.91]) (Latimier et al., 2021). This

section will first briefly cover how experiments of the testing effect with mathematical content can be designed. Then I will discuss potential pitfalls surrounding retrieval in experiments with mathematical content and what typical paradigms for spaced retrieval practice look like in the domain of mathematics learning. Finally, this section will finish with a review of potential moderators of spaced retrieval practice.

1.4.1 Typical Paradigms in Mathematics Spaced Practice Experiments

The literature which underpins this project falls broadly into three categories, pure spacing, pure retrieval practice, and spaced retrieval practice. Firstly, it is important to highlight a major difference between other domains and Mathematics. It is much more common to see across-session designs (Küpper-Tetzel, 2014). Across-session spacing (and spaced retrieval) experiments typically either focus on applying spacing and retrieval to one or two procedures and vary the inter-session interval and retrieval interval (Rohrer & Pashler, 2007; Rohrer & Taylor, 2006), or they apply the techniques to a cohort in a course, where the retrieval interval may differ for items presented at the start of the course and those at the end (Bego et al., 2017; Crissinger, 2015; Lyle et al., 2020). While lacking ecological validity, the isolated inter-session interval or retrieval interval experiments provide a purer measurement of the spacing effect than within course experiments. This is because as rearranging the material during the course inevitably leads to interleaving effects as well. Previous studies have attempted to isolate the spacing effect and interleaving effect. This is important as while all interleaving is subject to spacing, there is also an additional benefit to participants' ability to discriminate between problems (Chen & Kalyuga, 2021).

Early studies showed the important relationship between retrieval interval and inter-session interval and that the benefits of spacing are typically observed over longer periods of

time and can often be detrimental to immediate performance. For example, participants either massed ten combinatorics problems in one session or spaced them across two sessions with an inter-session interval of one week (Rohrer & Taylor, 2006). After one week, when tested, there was no significant effect of spacing, but there was a substantial large effect when tested four weeks later. However, in a similar experiment, one week was sufficient to see a large spacing effect (Rohrer & Pashler, 2007).

A second type of experimental design incorporates spaced retrieval practice into courses through cumulative testing. A common example would be to take a course currently running over a semester in a university and add weekly tests. For example, in Hopkins et al. (2016) the massed condition contained novel questions about that week's topic, while in the spaced condition there are typically some novel questions from the current topic, but the test will also consist of topics previously covered within the course. One reason this is good is that it is ecologically valid and easy to apply. On the other hand, it is more difficult to measure the specific effect of spacing on specific topics as items nearer the beginning may have a larger retrieval interval than those presented later. Furthermore, it is not possible to space materials learnt just before the exam. However, some have attempted to solve this problem by excluding topics learnt near the end of the course from their analyses (Hopkins et al., 2016; Lyle et al., 2020). Another potentially confounding factor is that simpler concepts presented at the start of the course may be used as building blocks for the later topics, which would mean these concepts get additional testing. Additionally, within this type of experimental design it is often difficult to see exactly how the individual items are spaced and how specifically they report the schedule varies from paper to paper. For example, some are vague where they say they change five to ten

percent of any homework assignment between the standard all novel homework versus the cumulative condition (Beagley & Capaldi, 2020).

1.4.2 Moderators of the Efficacy of Spaced Retrieval Practice

Other than the inter-session interval and retrieval interval mentioned earlier other factors are thought to modulate the efficacy of spaced retrieval. Individual differences may play a part in the efficacy of the spacing effect and learners' ability to implement a distributed practice routine. When students signed up for optional additional practice sessions for statistics, those in the distributed practice condition had a much higher rate of attrition than those on the massed practice schedule (Barzagar Nazari & Ebersbach, 2019). They also found that female students had a significantly higher chance of completing the practice sessions. However, that analysis was performed exploratorily, and additional confirmatory work is required to evidence their claims. As this was optional extra practice, this could have affected their study in two directions. First, they may have biassed their work towards those who want to work hard and have high conscientiousness. However, they may have also biassed the sample in the other direction, perhaps those who understood it all well didn't think they needed more practice, or parental pressure to sign up may have meant that only those who were already struggling in mathematics signed up. Complexity has previously been looked at in spacing and testing effect literature. A previous meta-analysis coded studies by overall complexity which "was defined by the degree to which the task requires a number of distinct behaviours, the number of choices involved in the performance of the task, and the degree of uncertainty involved in performance of the task" (Donovan & Radosevich, 1999). They found that overall complexity of the material was correlated with lower effect sizes.

1.5 Complexity in Mathematics

I chose to investigate task complexity as a moderator and potential boundary condition of the spacing effect for three reasons. Firstly, Donovan and Radosevich (1999) found that it was a significant moderator of the spacing effect in their meta-analysis. Secondly, the initial literature review did not find any studies that systematically investigate complexity and spacing, though Chen et al. (2024) was published during my PhD. And finally, if complexity is a mediating factor, then it may be easy to adjust algorithms or individualised learning systems that employ a spaced retrieval schedule, allowing the results of any research to have an immediate positive impact. Alternatively, if spacing is robust to changes in complexity then that provides further evidence to suggest its use in schools and across edtech platforms.

1.5.1 Defining Complexity

In the mathematics learning domain task complexity is often defined procedurally or conceptually (Crooks & Alibali, 2014). Previous reviews of the literature surrounding task complexity of mathematics material suggest that experiments that claim to measure conceptual complexity are often ill defined and not operationalized in theoretically relevant ways, finding that only 35% of the studies actually defined conceptual knowledge (Crooks & Alibali, 2014). Due to the lack of consensus on defining and operationalising conceptual complexity this project will initially focus on procedural complexity. While higher-level conceptual knowledge is important, basic procedural facts are still vital for students to gain proficiency in. For example, being able to quickly retrieve a procedure or basic fact is useful when solving more complex problems (Roediger & Pyc, 2012). Additionally, a procedure can be taught in isolation, while conceptual knowledge, which requires links between topics, cannot (Hiebert & Lefevre, 1986). This enables the selection of material, which requires few prerequisites and is novel to the

participants, which is useful experimentally. Within mathematics, education procedural knowledge is thought of as the knowledge of the steps of how to solve a particular problem. It is commonly measured by tracking participants' accuracy on problem solving tasks (Crooks & Alibali, 2014). Others further refine procedural knowledge in mathematics by separating out knowledge of the form, the formal language and symbols used to communicate mathematics, from knowledge of the rules, procedures, and algorithms to solve specific tasks (Hiebert & Lefevre, 1986). This is a useful distinction as it is important to ensure that the form of the mathematics does not impede participants ability to learn and retrieve the rule. For example, one spacing experiment left out the factorial symbol "!" when teaching participants a procedure (Rohrer & Pashler, 2007). This is an example of prioritising the algorithm required to solve the specific problem, while reducing the complexity of the form the problem is presented in.

In chapter 4 of this thesis, I switch from procedural complexity to element interactivity as the measurement of task complexity. Element interactivity was defined previously when discussing Working Memory Resource Depletion in the theories of the spacing effect section.

1.5.2 Complexity and Spacing

Several theories of the spacing effect may be affected by task complexity. The study-phase retrieval account, for example, could predict the effects of spacing would increase, decrease, or disappear entirely, depending on the experimental design. To illustrate this, imagine a hypothetical experiment where low or high complexity material is either spaced or massed. If the spacing condition were equal across complexity conditions, the experimenter could choose to either optimise the spacing of the low or high complexity material. The "optimal" inter-session intervals for low complexity material would mean they were retrieved just before they were forgotten (Cepeda et al., 2006), as this would require the most effort to retrieve them while still

successful resulting in the biggest boost to retrieval strength (Bjork et al., 2011). Assuming that more complex material is more difficult to retrieve given an equivalent amount of practice, then this inter-session interval would be sub-optimal for the more complex material, as many participants would not successfully retrieve the prior learning event, and participants would perform worse on a post-test. If instead the experiment was designed to optimise the spacing of the more complex material, then the lower complexity material would be retrieved more easily, requiring less effort, and therefore reducing the spacing effect.

It is difficult to predict whether contextual variability would be affected by the complexity of the material, as this theory relies on picking up contextual information around the learning session to provide future paths for retrieval. Perhaps if maximum attention is required to learn an item, then there is less opportunity to pick up other contextual cues.

The deficient processing theory may predict that low complexity material would be glanced over and not given sufficient processing therefore weakening the effect of spacing. Alternatively, more complex material may have sufficient time to be processed in the massed condition, but insufficient time to process in the spaced condition. This would predict that the relationship between complexity and spacing would be highly dependent on the scheduling conditions. If complexity reduces the efficacy of the spacing effect, then subsequent experiments could aim to find ways around this. If the spacing effect relies on successful retrieval, complexity may reduce the efficacy of spacing by reducing the chance of successful retrieval. In this case, shorter inter-session intervals may improve retention, as this will decrease the difficulty of retrieval.

1.6 Current Project

This literature review defined spacing and retrieval and looked at the factors that moderate the efficacy of spaced retrieval practice. Complexity appears to be an important moderators indicated in past meta-analyses (Donovan & Radosevich, 1999). Further, there is a lack of studies that systematically investigate complexity, and more research could be used to improve current spaced repetition algorithms. Therefore, this project will look at the following research questions: Does spaced retrieval practice work for mathematics learning? This question will be addressed through the meta-analysis in chapter 2. Does the procedural complexity of the material affect the efficacy of spaced retrieval practice? This question will be addressed through the first two experiments in chapter 3. Does element interactivity affect the efficacy of spaced practice? This question will be addressed in chapter 4.

2 A Meta-analytic Review of the Effectiveness of Spacing and Retrieval Practice for Mathematics Learning

Additional information for journal style thesis:

The following chapter has been reproduced with permission from Springer Nature after publication in Education Psychology Review.

Citation:

Murray, E., Horner, A. J., & Göbel, S. M. (2025). A Meta-analytic Review of the Effectiveness of Spacing and Retrieval Practice for Mathematics Learning. *Educational psychology review, 37*(3), 1-28. ***https://doi.org/10.1007/s10648-025-10035-1***

Statement of authorship:

This paper was authored by Ewan Murray and reviewed, edited, and supervised by Aidan J. Horner and Silke M. Göbel. Ewan Murray performed all data analysis, and the results were reviewed by Aidan J. Horner and Silke M. Göbel.

Data availability:

The data and code required to reproduce this manuscript, and all analyses is available on the OSF (***https://osf.io/qtfcu/***).

Abstract

Spaced retrieval practice harnesses two well-studied phenomena: the spacing effect, where spacing out practice over several sessions leads to a gain in retention compared to massed practice in one session; and the testing effect, where material that is tested is better retained than material that is restudied. This meta-analysis investigates if, and under what circumstances, spaced and retrieval practice can benefit mathematics learning. We found a robust small to medium effect of spaced versus massed practice overall (g = 0.28, 27 studies, 53 effect sizes). Those studies can be split into two subsets based on their experimental design, where material was either taught in isolation (10 studies, 27 effect sizes) or as part of a course (17 studies, 26 effect sizes). We found a larger, yet less robust, effect for the isolated learning (g = 0.43) than for course-embedded (g = 0.24). Our search also revealed 7 studies, 32 effect sizes, which manipulated testing versus restudy. The weighted mean effect of testing versus restudy was g = 0.18. However, the 95% confidence interval crossed zero, suggesting the testing effect is not robust. Overall, our results suggest that spaced practice can improve mathematics learning for material in isolation and within a course. However, the effect may be smaller than in other domains. Additionally, the current literature does not provide conclusive evidence for a consistent effect of retrieval practice for mathematics learning, possibly due to the smaller number of studies available.

Keywords: Spacing effect, Distributed practice, Testing effect, Retrieval practice, Mathematics

A Meta-analytic Review of the Effectiveness of Spacing and Retrieval Practice for Mathematics Learning.

Many decisions go into the design of a program of learning. What to learn and how to learn it are thought about in painstaking detail to maximise the chances of success for the learners. A critical question to ask when designing a learning program is when should revision of prior material take place? Precisely when a session takes place during a day is often out of the hands of instructional designers and up to timetablers, but once time has been allotted there is often freedom to choose what specific questions to ask and when. When an instructor decides which questions to ask during the start of a lesson, during the main tasks or as homework, it is here that there is the opportunity to improve learning through the use of retrieval practice and spaced repetition. There has been increased interest in harnessing retrieval practice, actively retrieving information rather than restudying it, and spaced practice, spreading out practice over multiple sessions rather than a single session, to improve mathematics learning, both in research and in applied settings. Given this recent spike in interest, it is valuable to synthesize the current research and see if these techniques are as effective for mathematics learning as they are in other domains. The goal of this meta-analysis is to review and synthesise the current literature surrounding the spacing and testing effect in mathematics learning.

2.1.1 Overview of Research on Spaced Practice

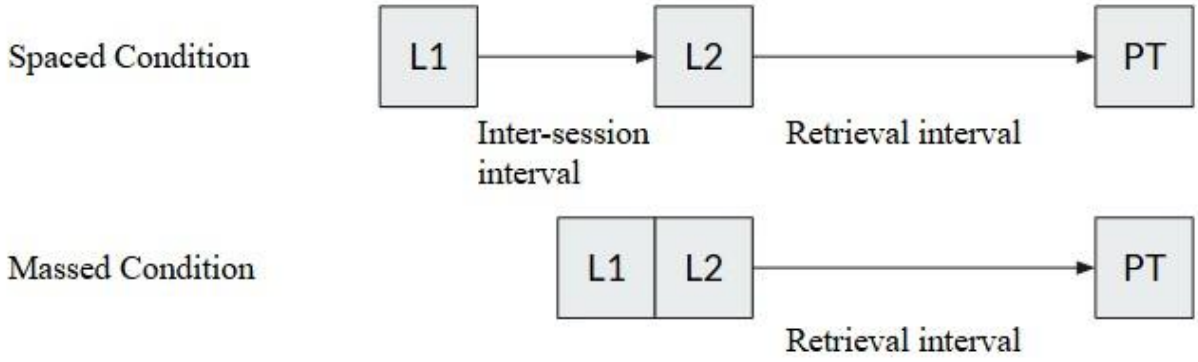
The spacing effect is a special case of the distributed practice effect, where the temporal spacing of practice or presentations of stimuli are varied and a change in retention is observed (Delaney et al., 2010). The spacing effect is found within session, by varying the time between presentations of a stimuli and across session, by varying the times between practice sessions (Küpper-Tetzel, 2014). All the spacing effect studies included in this meta-analysis are across-

session studies, so we will define the spacing effect as the change in retention when an equivalent amount of practice is spread over multiple sessions versus a single massed session (see Figure 1-1). Broad ranging meta-analyses across all domains found medium to large effects of spacing versus massed practice (g = 0.46, Donovan & Radosevich, 1999), with some more specific ones such as the domain of L2 Language learning finding larger effects (g = 0.80, with a delayed post-test, Kim & Webb, 2022). The effect has been observed across many different domains, age groups and even species (Delaney et al., 2010).

Many theories have been proposed to describe the mechanisms underlying the spacing effect (Delaney et al., 2010; Küpper-Tetzel, 2014; Maddox, 2016; for reviews, see Toppino & Gerbier, 2014). The three most discussed theories are *study-phase retrieval*, *deficient processing*, *encoding variability,* or combinations of those three (Toppino and Gerbier, 2014). Study-phase retrieval suggests that successful recall of the initial presentation of the to be recalled item results in a boost to future recall (Hintzman, 2004; Thios & D'Agostino, 1976). Deficient processing suggests that repeated presentations of an item in the massed condition leads to lower overall processing, when compared to items spaced out over time, which in turn leads to a lower quality of learning overall (Hintzman, 1974). Encoding variability theories suggest that over time the context around an item drifts and is encoded when presented to the participant, when presentations of the item are spaced out then this increases the probability of the retrieval context being more similar to one of the spaced encoding contexts relative to two the massed encoding presentations (Glenberg, 1979). These theories and reviews informed our choice of moderators (see supplementary material), along with discussions of relevant moderators in recent studies and reviews (Emeny et al., 2021; Latimier et al., 2021). However, as our moderation analyses are unable to help distinguish between these theories we do not go into further detail.

Figure 2-1

Simple Spacing Effect Paradigm



Note. A diagram outlining the overall procedure of a hypothetical spaced vs massed learning experiment. L = Learning session, PT = Post-test.

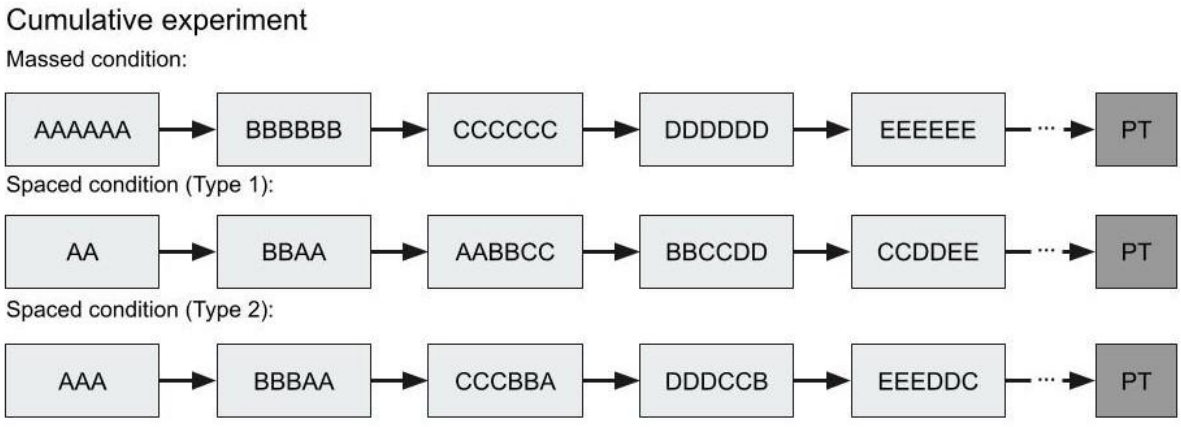
Within mathematics learning, spacing effect studies follow one of two experimental designs based on how the material is integrated into the larger educational environment: isolated learning versus material incorporated into a course. Some studies that use mathematics material require learners to practice isolated learning outside of the confines of a course. In these cases, the participants are taught a set number of mathematical procedures or are given practice solving problems in such a way that there is a clear inter-session interval and retrieval interval (Rohrer & Pashler, 2007; Rohrer & Taylor, 2006) and follow a similar experimental design to previous verbal learning studies (see Figure 2-1).

Another common structure for these studies is to embed the spacing schedule into a course already taking place. Suppose there are five learning objectives and six questions for each objective (see Figure 2-2). There are various ways in which these practice questions can be presented to learners. One way is to present them in a massed manner where after learning about

the topic the learners are presented with all available questions on that topic. This is how most mathematics textbooks are laid out (Rohrer et al., 2020), with few notable exceptions (Hake, 2012). On the other hand, they could be presented uniformly spaced across multiple sessions (see Figure 2-2, spaced condition type 1). Examples of this type include Lyle et al. (2020) and Hopkins et al. (2016). Alternatively, it is possible to start with more practice examples in the first instance and reduce the number of items in subsequent practice sessions; this type of spacing was used by (Holdan, 1986) (see Figure 2-2, spacing condition type 2). Others allow for variations in the exact spacing and amount of practice by implementing a formula for teachers to use (Hirsch et al., 1982). However, these course-embedded studies provide a less pure measure of spacing as interleaving, mixing practice of different topics and skills, is also involved.

Figure 2-2

Experimental Design for Course-embedded Spacing Study



Note. A diagram outlining potential procedures for course-embedded spacing experiments.

It has been noted that increased *contextual interference* or *interleaving*, that is, shuffling the order of practice problems from tasks to increase the variability of practice, can lead to greater retention and transfer of skills versus blocked practice (Shea & Morgan, 1979). We do

not include studies that purely investigated blocked versus interleaved practice. However, we did not exclude distributed practice studies which, due to overlapping practice schedules, interleaved practice material. In the included studies, the manipulation focused on the practice schedule of each item not carefully manipulating interleaving versus blocked practice throughout. We believed it was important to include these studies as this is likely how spacing would be implemented in a classroom. Previous mathematics learning interleaving experiments comparing interleaved with blocked practice have shown that alongside the benefit of spacing there is an additional benefit of discriminative contrast (Foster et al., 2019). A meta-analysis found a small to medium effect ($g = 0.34$) of interleaved versus blocked practice for mathematics learning materials (Brunmair & Richter, 2019). In an educational setting where the aim is to maximise learning with considerable time constraints, this is the way spacing would be implemented with old material intermixed with the new. Therefore, even though there is an effect of interleaving, experiments that follow this kind of structure were included, as all the studies claimed to be harnessing the spacing effect. As this is likely to be how spacing would be implemented in an actual educational setting, these studies bring increased ecological validity.

*2.1.2 Ov*erview of Research on Ret*rieval Practice*

The testing effect, or retrieval practice, is a boost in retention to material that has to be retrieved from memory compared to material that is restudied. Retrieval practice has been shown to be a powerful way to improve retention (Roediger & Butler, 2011), but there has been little research into its efficacy with mathematics. Multiple previous meta-analyses have investigated the testing effect in other domains. Rowland (2014) ($g = 0.50$, 95% CI [0.42, 0.58]), Adesope et al. (2017) ($g = 0.61$) and Yang et al. (2021) ($g = 0.50$, 95% CI [0.442, 0.557]) all found a medium to large boost to retention when retrieval practice was used. Yang et al. (2021) included

27 effect sizes from 14 studies that pertained to mathematics or statistics material, however, of those, only two studies compared testing to restudying and were included in our sample (Betsch et al., 2015; Dirkx et al., 2014). The other control conditions included in Yang et al. (2021) included more versus fewer questions, elaborative strategies, or no activity (just a distractor task after the initial study session), which were outside the scope of our pre-registration. There is also meta-analytic evidence that retrieval practice can enable transfer in a variety of contexts such as different test formats and problem types (Pan & Rickard, 2018), but perhaps not to untested information studied during the initial practice. However, there has been no meta-analysis that focused specifically on retrieval practice in mathematics versus restudy.

It is not immediately apparent what restudy would look like in mathematics, in comparison to verbal learning studies where a word pair could be presented again. Indeed, later in this meta-analysis we discuss the difficulty in understanding precisely how much students had to retrieve information, and when they were able to fall back on their notes. One way to manipulate retrieval versus restudy in mathematics is to compare completing practice where retrieval is necessary to worked example problem pairs (Fazio, 2018). Worked examples can be effective learning procedures as they allow learners to focus on the task at hand without the need to hold all parts of the problem in working memory at once (Sweller, 2006). For example, Fazio (2018) compared retrieval practice with worked example problem pairs. In the retrieval condition they were tested on the procedure, but in the worked example condition they were presented with a worked example they could follow along with, so they may not have had to retrieve the procedure. They found that worked examples increased performance on an immediate test, but the retrieval condition produced greater retention on a delayed test. Alternatively, Dirkx et al. (2014) compared reading a text on how to calculate a probability, versus alternating between

reading and testing. They found that reading and testing led to both better fact recall and a better ability to apply the procedure.

Furthermore, spacing and testing are often combined into *spaced retrieval practice*. Latimier et al. (2021)'s previous meta-analysis suggested that the effect of spacing and testing is additive and is typically a large effect (g = 1.01, 95% CI [0.68, 1.34]).

2.1.3 The Present Meta-analysis

This meta-analysis focuses on studies that involve the learning of mathematics. We attempted to answer three key questions. First, does the spacing of mathematics practice lead to higher retention (versus massed)? Second, does retrieval practice of mathematics lead to higher retention (versus restudy)? Third, we planned to investigate whether the combination of spacing and retrieval lead to higher retention of mathematics knowledge (versus massed retrieval), however, this was not possible with the available studies. Additionally, we investigated whether any of the potential moderating variables (such as the length of the retrieval interval or type of material being studied) have any effect on the heterogeneity or mean effect sizes, however, as these exploratory analyses provided little informative information they are included only in the supplementary material.

Based on previous meta-analyses of spacing and retrieval practice with non-mathematical material we hypothesized that we would find a robust effect of testing and spacing for mathematics material and that the combination of spacing and retrieval would be additive.

2.2 Method

2.2.1 Selection Criteria

A librarian, with subject area expertise, was consulted to develop the Boolean search queries, as suggested in recent meta-analysis recommendations (Hansen et al., 2022; Steel et al.,

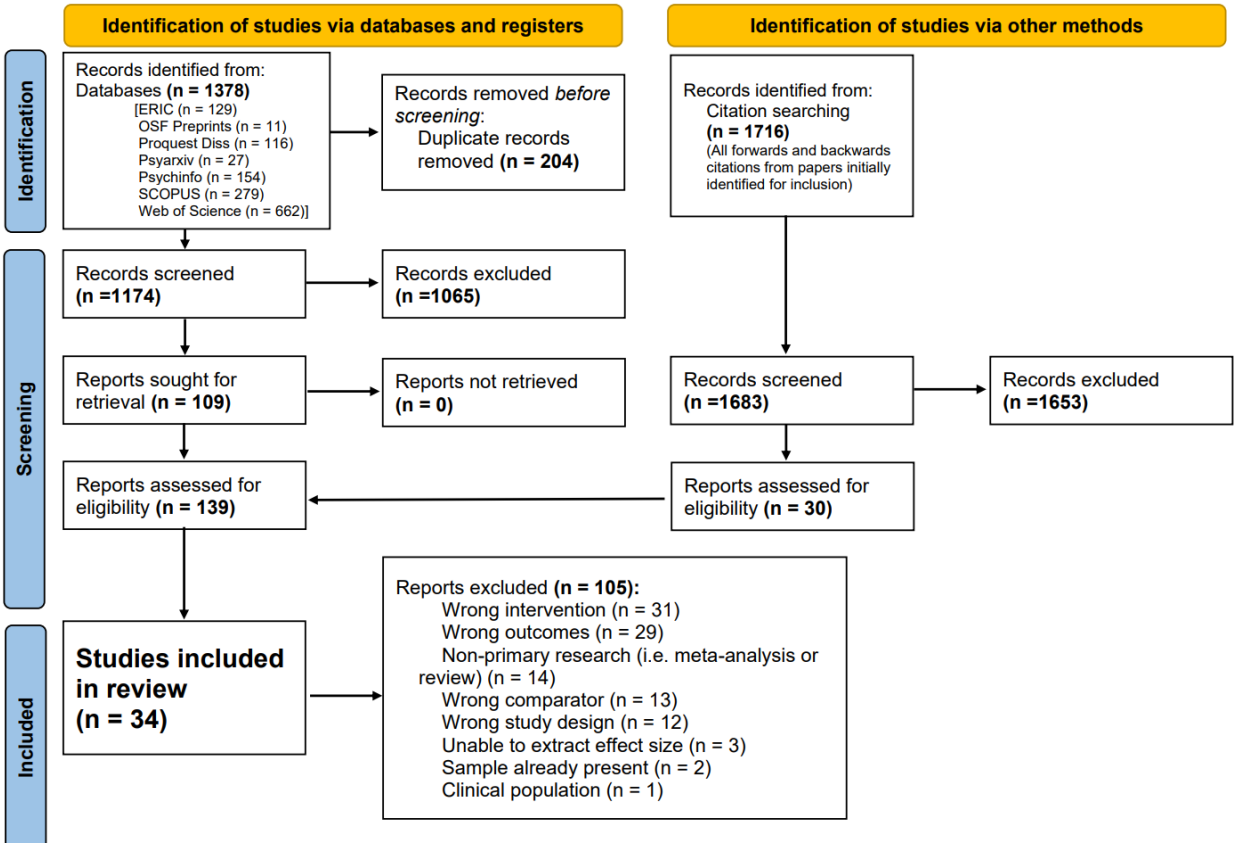
2021). We first looked for any material related to distributed or retrieval practice (using a variety of terms) and then narrowed that down to only studies that mentioned mathematics. The exact query for each database has been made available to increase reproducibility in the supplementary material. At this stage, any article that mentioned the use of an intervention for mathematics education based on the spacing effect or retrieval effect was included. The database search was undertaken on the 11th of January 2023 and then updated after initial manuscript review on 25th of February 2025.

There are several papers that were excluded during the screening phase, that at first glance may have seemed applicable. We excluded three papers that used explicit timing designs to incorporate distributed practice because of their outcome measure: digits correct per minute (DCPM) (Bullard, 2020; Powell, 2022; Schutte et al., 2015). Digits correct per minute (a measure of fluency), did not seem comparable to a mean percentage score on a post-test (a measure of retention) which was the outcome specified in our pre-registration (***https://osf.io/qtfcu/***). An additional three papers were excluded where the mean percentage score, standard deviation or number of participants wasn't provided and we were thus unable to extract them (Rea & Modigliani, 1985; Reed, 1924; Yazdani & Zebrowski, 2006), in each case we were not able to contact the authors.

2.2.2 Data Collection

Figure 2-3

Prisma Flow Diagram



Note. Diagram outlining how many papers were identified and from where and how they came to be included or excluded in the final sample of 34 papers.

Alongside the database search a request for information was posted in the Mathematical Cognition and Learning Society mailing list but generated no replies. Additionally, authors included after full screening were asked via email whether they had any unreported data, to combat the file drawer problem (Rosenthal, 1979). Excluding those authors for whom no email was available, all but three authors responded to this question, however, no one reported any unpublished data.

After searching the databases, the title and abstract screening phase and full text reviews took place on *Covidence Systematic Review Software* (2023), this information is presented in the form of a PRISMA flowchart (see Figure 2-3) Two screeners then screened ten percent of titles and abstracts (n = 88) to make sure that the inclusion and exclusion criteria (see Table 2-1) were being applied reliably and worked (88.6% inter-screener agreement rate). The inclusion and exclusion criteria were informed by the PICOS framework (see Appendix). One screener then screened the remainder. Texts were then sent to full text review. Two screeners fully screened 10 of the papers (100% inter-screener agreement rate) before the remainder were screened by a single screener. The information required to estimate mean effect sizes, moderation and quality analyses was then extracted into an excel file containing the extraction table. During the extraction phase, the extraction table was evaluated by having two researchers extract the same five randomly selected papers, then meet up to discuss any discrepancies between the two sheets. This led to our terminology being clarified and we edited the extraction sheet to better fit the format of the studies being captured. A further three papers were extracted as a pair to check the updated extraction sheet was appropriate. Then a single screener extracted the remainder of the papers. The full data extraction sheet is available in the OSF repository (***https://osf.io/qtfcu/***). After the initial extraction nineteen authors were contacted regarding missing information. Nine replied (47%). Only one paper was unable to be included as critical information related to the calculation of effect sizes was unavailable and no reply was received. For six of the papers published before 2000, no current contact information for the researcher was found, but this did not prevent their inclusion.

Table 2-1

Inclusion and exclusion criteria

Inclusion	Exclusion
The sample use either spaced repetition or retrieval practice (or both)	Non-experimental designs
Participants required to learn mathematics material	Clinical populations
Performance measured on a post-test	Effect size based solely on the mathematics material unable to be extracted
English abstract	

Note. These inclusion and exclusion criteria were visible to screeners during all phases of data collection.

2.2.3 Computation of Effect Sizes

In all the studies included in this meta-analysis the recorded effect compares the difference in performance on a post-test between two groups. In the first analysis we have included the studies comparing spaced versus massed practice and in the second the studies which compared testing versus restudying.

The escalc function in the R module metafor (Viechtbauer, 2010) was used to calculate Hedges (1981)'s g for each extracted effect. As most studies did not report the correlation for within-subject effects $r = .5$ was used to compute the pooled standard error as used in previous

spacing and retrieval meta-analyses (Latimier et al., 2021; Rowland, 2014). For between and within participant effects the d was calculated as follows:

$$(1) \quad d = (m_1 - m_2)/S$$

with the pooled standard deviation S for between subject effects calculated as:

$$(2) \quad S = \sqrt{\frac{(n_1 - 1)sd_1^2 + (n_2 - 1)sd_2^2}{n_1 + n_2 - 2}}$$

Where n_1 and n_2 are the number of participants in the first and second group, m_1 and m_2 are the observed means the first and second group and sd_1 and sd_2 are the observed standard deviations of the first and second group. The sample variance for between subjects was calculated using the following formula (Hedges, 1981):

$$(3) \quad V = \frac{n_1 + n_1}{n_1 * n_1} + \frac{d^2}{2(n_1 + n_1)}$$

and the standard error, $SE = \sqrt{V}$.

For within participant effects the pooled standard deviation S was using raw score standardization (4). The variance for within-subject effects was calculated as follows (Becker, 1988; Viechtbauer, 2010) (5).

$$(4) \quad S = \sqrt{\frac{sd_1^2 + sd_2^2}{2}} \qquad (5) \quad V = \frac{\frac{1(1 - r)}{n} + d^2}{2n}$$

Then for each subset of studies the overall weighted mean effect size was calculated using the metafor and clubSandwich modules in R. Using the guidelines provided by Pustejovsky and Tipton (2022) we chose to perform a Robust Variance estimation with correlated and hierarchical effects (CHE). We chose this method as many of the studies contribute multiple effect sizes but also often aim to find boundary conditions for the spacing or

testing effect therefore within-study heterogeneity is expected. Using this method allows us to account for the correlation between effects from the same experiment and the hierarchical structure resulting from experiments reporting multiple effects. Heterogeneity is a measure of how different the effects are from one another, if there is a true effect with little variance then you would predict little heterogeneity.

Due to the small sample size, the Q statistic would not have been an appropriate measure of heterogeneity (Gavaghan et al., 2000). Instead, we calculated I^2 using the dmetar package (Harrer et al., 2019), which uses the formula outlined in Cheung (2014) to calculate I^2 for each level of the hierarchical meta-analysis. In some cases, the initial analyses suggested there was very little heterogeneity, however visual inspection of the forest plots suggested this was unlikely. As I^2 can still be highly uncertain for between-study heterogeneity when sample size is small (Chung et al., 2013), confidence intervals were calculated to measure this uncertainty (Viechtbauer, 2017).

2.2.4 Risk of Bias

To check for the risk of publication bias we performed Egger's regression tests (Egger et al., 1997) on univariate versions of the meta-analyses. We also provide a corresponding funnel plot for each analysis. Additionally, we coded each source with quality indicators. The coding sheet was a modified version of Hjetland et al. (2020)'s coding sheet. We coded each extracted effect for sampling procedure used (0 - random, 1 - convenience), whether the test reliability was reported (0 - reported, 1 - not reported), if there were any floor or ceiling effects present (0 - not present, 1 - present), how missing data was dealt with (0 - better than list-wise deletion, 1 - list wise deletion), statistical power/sample size (0 - [N ≥ 150], 1 - [70 ≤ N <150], 2 - [N < 70] participants), whether attrition was reported (0 - reported, 1 - unreported). The scores were

summed, and the overall quality score was used as a moderator to see if it accounted for a significant amount of heterogeneity.

Outliers were checked using the metafor R package (Viechtbauer, 2010) to calculate the Cook's distance and difference in fit (DFFITS) values. The difference in fit value shows how many standard deviations the calculated mean effect size changes when a particular study is removed (Viechtbauer & Cheung, 2010).

2.2.5 Potential Moderators

Due to small sample sizes and overly prevalent categories, we diverted from our preregistered method of coding multivariate moderators to, whenever possible, uni-variate form. This was because when the adjusted degrees of freedom (Satterthwaite, 1946) are below four, the moderator analyses tend to over-reject and are therefore unreliable (Tipton, 2015). For example, Latimier et al. (2021) also changed from multivariate to univariate due to small sample sizes. In the extraction sheet we have left the original multivariate categories, and the coding process is explained in full in the pre-registration. Despite this change none of the results were significant. As our moderating analyses were exploratory, we do not include them in this manuscript and instead they are available as a supplementary analysis.

2.3 Results

This meta-analysis considers 34 studies with a total of 85 effect sizes. During the search and data collection processes it became clear that there were different ways that studies involving spacing could be broken down into subsets (isolated learning vs embedded into a course). We begin with an overview of all spacing studies before running further analyses on these two subsets. We then analyse retrieval studies that do not manipulate spacing.

2.3.1 Spacing Effect in Mathematics Learning

Our first pre-registered analysis tested the effect of spaced versus massed practice in mathematics learning. The systematic review revealed 27 studies with 53 effect sizes. We found a weighted mean effect of $g = 0.282$ (se = 0.045) (95% confidence interval [0.188, 0.376]). However, upon further review it became clear that this analysis consisted of two different experimental designs. We first review studies that focus on learning a single, or small number of, mathematical skill(s) in isolation where all items have the same retrieval interval. We will refer to these studies as *isolated learning studies*. In contrast, we then review studies that involve spaced versus massed practice integrated into a course. We will refer to these studies as *course-embedded studies*. In course-embedded studies material covered at the beginning of the course has a longer retrieval interval than material covered at the end of the course and simpler material may also be incorporated into more complex material further along the course. Running a categorical meta-regression between isolated learning studies and course-embedded studies found that the effects were larger on average for isolated learning than course-embedded studies (see below), although they do not differ significantly ($\beta = 0.188$, $SE = 0.106$, $p = .0952$). However, as we believe they are fundamentally measuring different effects, a pure measure of spaced practice (isolated learning) versus spaced and interleaved practice (learning embedded in a course), we ran all further analyses separately for the two subsets.

2.3.2 Sp*acing Effect in* Mathematics Learning - Isolated Lear*ning*

This subset consists of 10 studies with 27 effect sizes. Several studies focused on learning a simple combinatorial procedure (Ebersbach & Barzagar Nazari, 2020b; Emeny et al., 2021; Rohrer & Pashler, 2007; Rohrer & Taylor, 2006). Other areas of mathematics studied include

arithmetic (Barzagar Nazari & Ebersbach, 2019; Chen et al., 2018), algebra (Chen et al., 2018; Gay, 1973) and statistics (Ebersbach & Barzagar Nazari, 2020a).

All the isolated learning studies were peer-reviewed articles. For the country moderator most of the studies were sampled from the USA, with others from Germany (Barzagar Nazari & Ebersbach, 2019; Betsch et al., 2015; Ebersbach & Barzagar Nazari, 2020a), the UK (Emeny et al., 2021) and China (Chen et al., 2018). Many of observed effects took place in a pre-university setting, for example, Emeny et al. (2021) in a secondary school and Chen et al. (2018) in a primary school. The mean age of participants ranged from 9.23 years (Chen et al., 2018) to 24.42 years (Ebersbach & Barzagar Nazari, 2020a).

The only study to use a within-subject design was Emeny et al. (2021). Retrieval intervals ranged from a single day (Chen et al., 2018) to six weeks (Barzagar Nazari & Ebersbach, 2019). The most common retrieval interval was one week. Inter-session intervals ranged from a single day (Chen et al., 2018) to two weeks (Gay, 1973). The most common inter-session interval was one day. Over half of the measured effects (61%) provided corrective feedback, the remainder provided no feedback. Under half of the studies had the first retrieval session immediately after the first-time participants learned the item, while the remainder had a delay. Each study had either two or three sessions in total. Mean performance at the first time point was on average 61% across all observations but ranged from 30% (Ebersbach & Barzagar Nazari, 2020a) to 95% (Rohrer & Pashler, 2007). The number of items (procedures, facts) participants had to learn, where reported, ranged from one to eight. The number of times the learning objective was exposed to participants ranged from three times to twelve (Emeny et al., 2021).

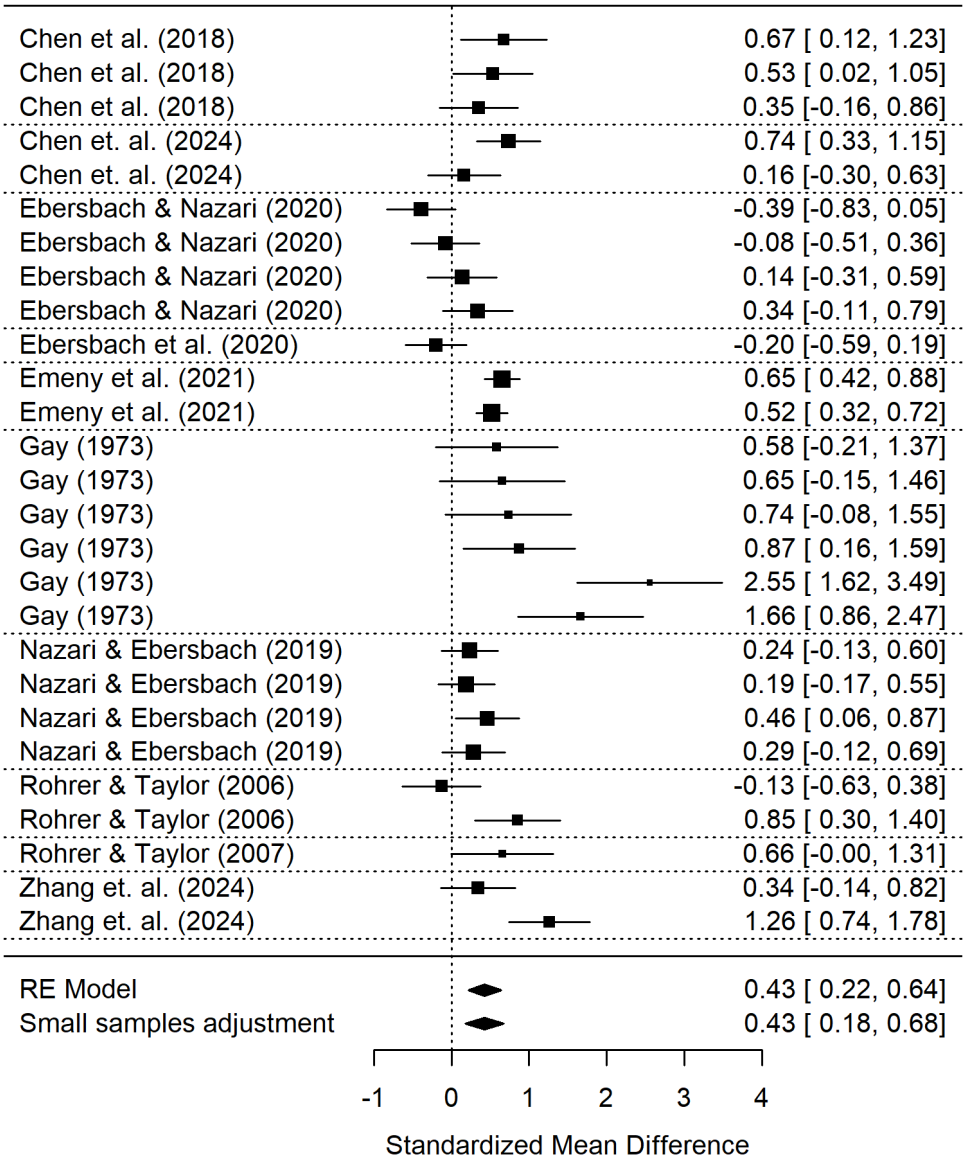
Overall Effect Size. The overall weighted mean effect size of spaced versus massed practice for isolated learning is $g = 0.427$ (se = 0.107) (95% confidence interval [0.179, 0.675]) (see Figure

2-4). The hierarchical structure of the meta-analyses can allow us to see how much variance was associated with each level of the hierarchy. Firstly, I^2 = 24% of the variance was associated with the first level (sampling error), I^2 = 33% was associated with the second level (within study heterogeneity) (95% confidence interval [1.456, 82.675]) and I^2 = 43.287% associated with the third level (between-study heterogeneity) (95% confidence interval 0, 88.234]).

When the Cook's distances for each effect were calculated none were greater than 0.5, suggesting that no single study was likely to be highly influential. In contrast, the difference in fit analysis shows that there are two dominant studies whose removal shifts the mean effect size by over half a standard deviation. Those studies are, firstly, Ebersbach and Barzagar Nazari (2020a) ($DFFITS$ = -0.708), which changes the mean effect size to g = 0.499 (se = 0.097) (95% confidence interval [0.268, 0.731]) and, secondly, Gay (1973) ($DFFITS$ = 0.638) whose exclusion changes the mean effect size to g = 0.368 (se = 0.096) (95% confidence interval [0.14, 0.596]). Critically, 95% confidence intervals remain greater than zero when either of these studies are removed, demonstrating the robustness of the effect.

Figure 2-4

Isolated Learning Studies - Forest Plot



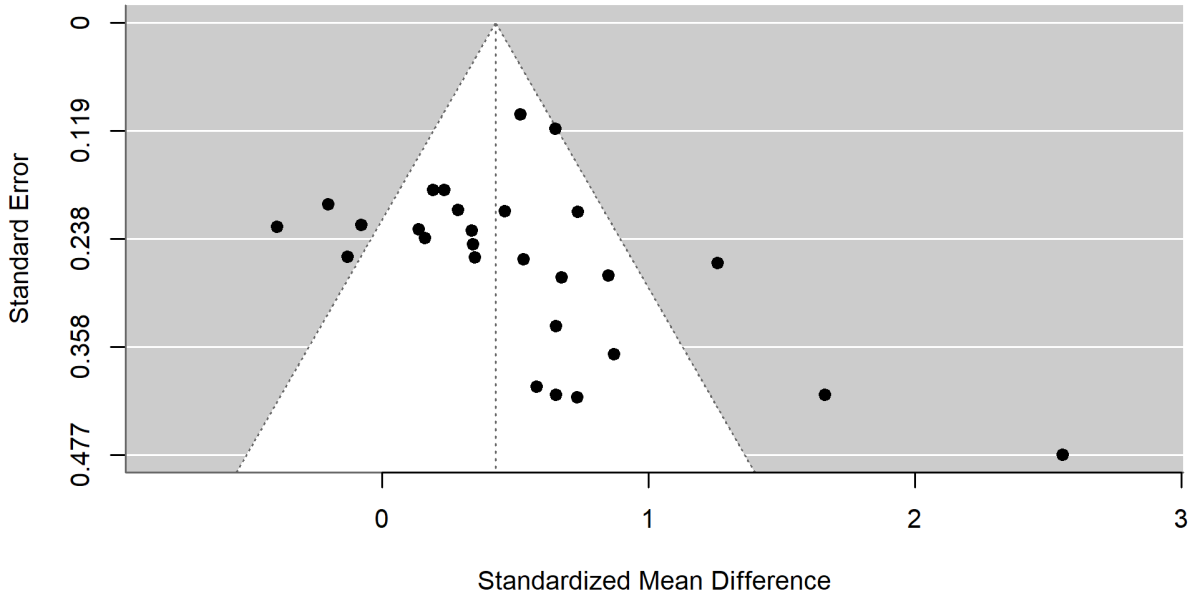
Note. A forest plot displaying the weighted mean effect sizes for each effect in the subset. The size of the square is proportional to the sample size; the error bars represent the 95% confidence

interval. The diamond at the bottom represents the overall weighted mean effect size before and after the small sample adjustment.

Risk of Bias. An insignificant Egger's regression test for funnel plot asymmetry ($t = 1.033$, d.f. = 25, $p = 0.312$) does not suggest publication bias resulting from non-significant results remaining unpublished. In contrast, the funnel plot (see Figure 2-5) appears to show asymmetry. There appears to be a significant gap of lower powered observed spacing effects that were smaller or negative, which could be an indicator of publication bias. This is balanced by effects with a smaller standard error that were closer to zero or negative, which is why the asymmetry may not appear in the Egger's regression test. Furthermore, when a trim and fill analysis was applied, it suggested there were no studies that needed to be added.

Figure 2-5

Isolated Learning Studies - Funnel Plot



Note. A funnel plot showing the effect size of spaced practice versus massed practice against the standard error for each effect in the sample. Large asymmetry would suggest publication bias.

Overall, these analyses suggest that there is a medium to large beneficial effect of spaced versus massed practice when a particular mathematical procedure/skill is taught in isolation under more controlled conditions.

2.3.3 Spaced versus massed practice - Course-embedded Studies

This subset consists of 17 studies with 26 effect sizes. A variety of mathematical areas are used for the material including algebra (Camp, 1973; Goettl et al., 1996; Holdan, 1986; Lerma, 1990; Reed, 1924), arithmetic (Moss, 1996; Weaver, 1976), calculus (Beagley & Capaldi, 2016, 2020; Bego et al., 2017; Gorgievski, 2012; Lyle et al., 2020, 2022) and statistics (Crissinger, 2015).

All studies recruited participants in the USA. More than half (60%) used varying inter-session intervals based on a formula (e.g., Hirsch et al., 1982) to distribute questions on a particular topic or skill, while others had an inter-session interval for each learning objective that remained constant (Lyle et al., 2020). Three quarters of the studies provided some form of feedback while a quarter did not. Goettl et al. (1996) was the only study that was run in a lab setting, but was still structured as a course, while the remainder were embedded into a course in a classroom or as homework. The mean age of participants was only reported for 23% of observed effects, but for those effects the mean age was 17.4 years old.

Over half the effects were extracted from peer reviewed articles (54%), while the remainder were extracted from theses (i.e., Camp, 1973; Gorgievski, 2012) or a conference paper (Bego et al., 2017). Many of the studies took place in a university setting (65% of effects) while the remainder either took place in secondary schools or were drawn from an adult population not in University (Goettl et al., 1996). Studies that took place in the classroom or used untimed homework could have allowed pupils to look back at their work, making retrieval uncertain (53% of observed effects) while studies that used timed quizzes in class or at home reduced that chance and were coded as requiring retrieval. Most of the studies adopted an expanding inter-session interval design (68% of observations) while the remainder used a uniform design. Thirty percent of effects were within subjects while 70% were between subjects.

The number of items refers to the number of learning objectives that were intended to be learnt and tested at the final exam, this ranged from seven (Reed, 1983) to forty-eight (Hopkins et al., 2016) where reported. The number of exposures refers to how many times each learning objective is practiced and ranged from three (Hopkins et al., 2016) to nine times (Gorgievski, 2012). Mean performance at first time point ranged from 11% (Goettl et al., 1996) to 85%

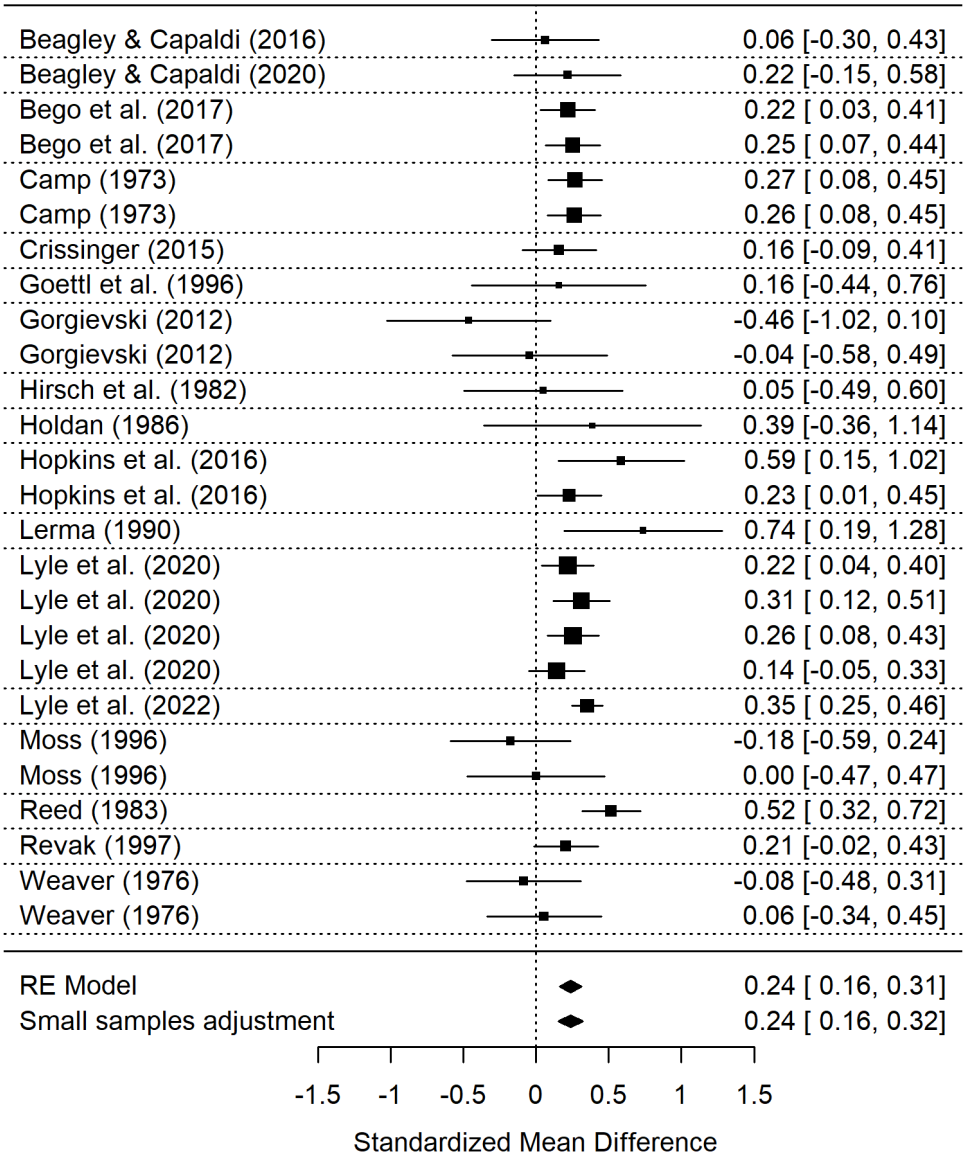
(Hirsch et al., 1982). Mean inter-session interval length ranged from one day (Goettl et al., 1996) to two weeks (Lyle et al., 2020). Time from final practice/retrieval session to exam ranged from two days (Camp, 1973) to five weeks (Hopkins et al., 2016; Lyle et al., 2020; Weaver, 1976) and five weeks was the most common delay.

Overall Effect Size. The overall weighted mean effect size of spaced versus massed practice for course-embedded material is g = 0.24 (se 0.038) (95% confidence interval [0.155, 0.324]) (see Figure 2-6). The hierarchical structure of the meta-analyses can allow us to see how much variance was associated with each level of the hierarchy. Firstly, I^2 = 57% of the variance was associated with the first level (sampling error), I^2 = 0% was associated with the second level (within study heterogeneity) (95% confidence interval [0, 50.555]). However, there was a large amount of uncertainty in the between-study heterogeneity due to the small sample size (Chung et al., 2021). Finally, I^2 = 43.065% associated with the third level (between-study heterogeneity) (95% confidence interval [0, 78.745]).

Again, the Cook's distances for each effect were calculated and none were greater than 0.5 which suggested that no single study was likely to be highly influential. The difference in fit analysis shows that study 15 (Lyle et al., 2022) shifts the mean effect size just under half a standard deviation ($DFFITS = 0.491$), the removal of this study changes the mean effect size to g = 0.24 (se = 0.038) (95% confidence interval [0.155, 0.324]). The studies that produced the next largest difference in fit were Hirsch et al. (1982) ($DFFITS = 0.638$) and Gorgievski (2012) ($DFFITS = 0.147$). The removal of Hirsch et al. (1982) changes the mean effect size to g = 0.256 (se = 0.036) (95% confidence interval [0.176, 0.337]).

Figure 2-6

Course-embedded Studies - Forest Plot



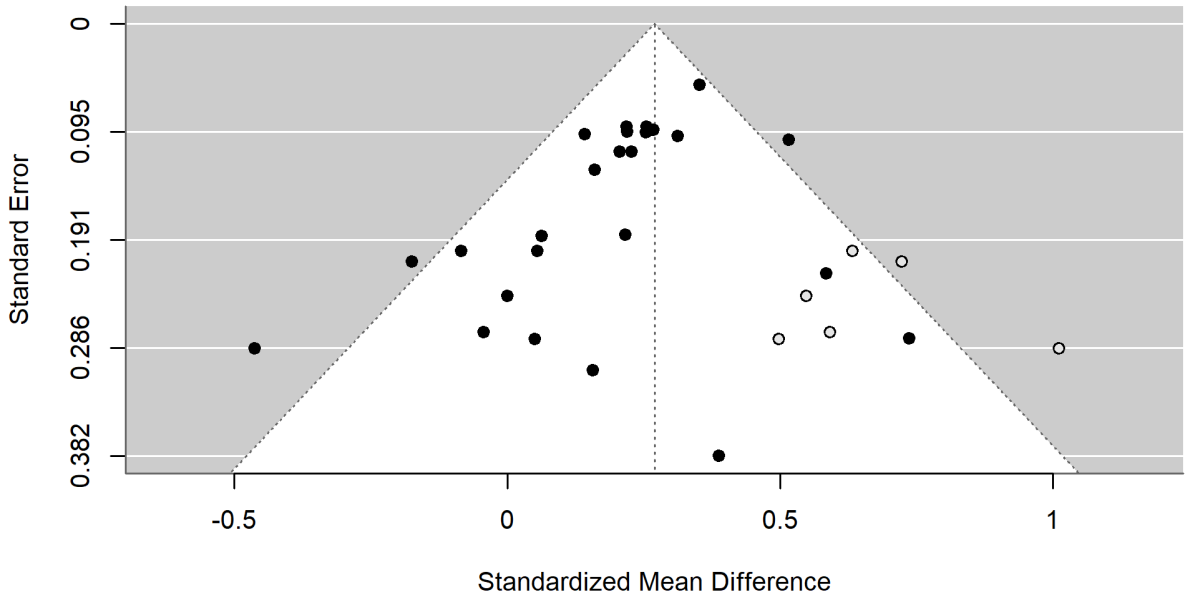
Note. A forest plot displaying the weighted mean effect sizes for each effect in the subset. The size of the square is proportional to the sample size; the error bars represent the 95% confidence

interval. The diamond at the bottom represents the overall weighted mean effect size before and after the small sample adjustment.

Risk of Bias. Egger's regression test for funnel plot asymmetry is significant (t = -2.4, d.f. = 24, p = 0.025) suggesting publication bias resulting from insignificant results remaining unpublished. The negative t-value would suggest under-reporting of results that exceed the mean effect size. The funnel plot (see Figure 2-*7*) does not show obvious asymmetry. When the trim and fill method (Duval & Tweedie, 2000) is applied to the uni-variate version of the model it adds in 6 studies and provides a corrected effect of $g = 0.27[0.215, 0.326]$.

Figure 2-*7*

Course-embedded Studies - Funnel Plot



Note. A funnel plot showing the effect size of spaced practice versus massed (for course-embedded studies) against the standard error for each effect in the sample. Large asymmetry

would suggest publication bias. The black circles represent the studies included in the main analysis, while grey studies are studies suspected to be missing by the trim and fill method.

In summary, there is a small to medium positive effect of spaced versus massed practice when the material is incorporated into a course structure, it appears to be a numerically smaller effect than for isolated learning.

2.3.4 *Retrieval versus Restudying in Mathematic*s Learning

This subset consists of 7 studies with 32 effect sizes. Fazio (2018) focused on learning and applying a particular procedure (fraction multiplication), while the others used a mixture of learning procedures alongside concepts or using the knowledge in problem solving tasks. The most common topic was statistics (Eustace et al., 2020; Lyle & Crawford, 2011; Yeo & Fazio, 2019), with the other studies focusing on probability (Dirkx et al., 2014), geometry (Betsch et al., 2015) or arithmetic (Fazio, 2018).

Half of the studies were undertaken in the USA, with the other three sampling from populations in the Netherlands (Dirkx et al., 2014), Germany (Betsch et al., 2015) and Ireland (Eustace et al., 2020). All studies took place in a university setting except one in a Secondary setting (Dirkx et al., 2014) and one in a primary setting (Betsch et al., 2015). Where reported, the mean age of participants ranged from approximately nine (Betsch et al., 2015) to twenty-two years old (Lyle & Crawford, 2011). All the included studies were peer-reviewed articles except for Fazio (2018) which was published as a book chapter.

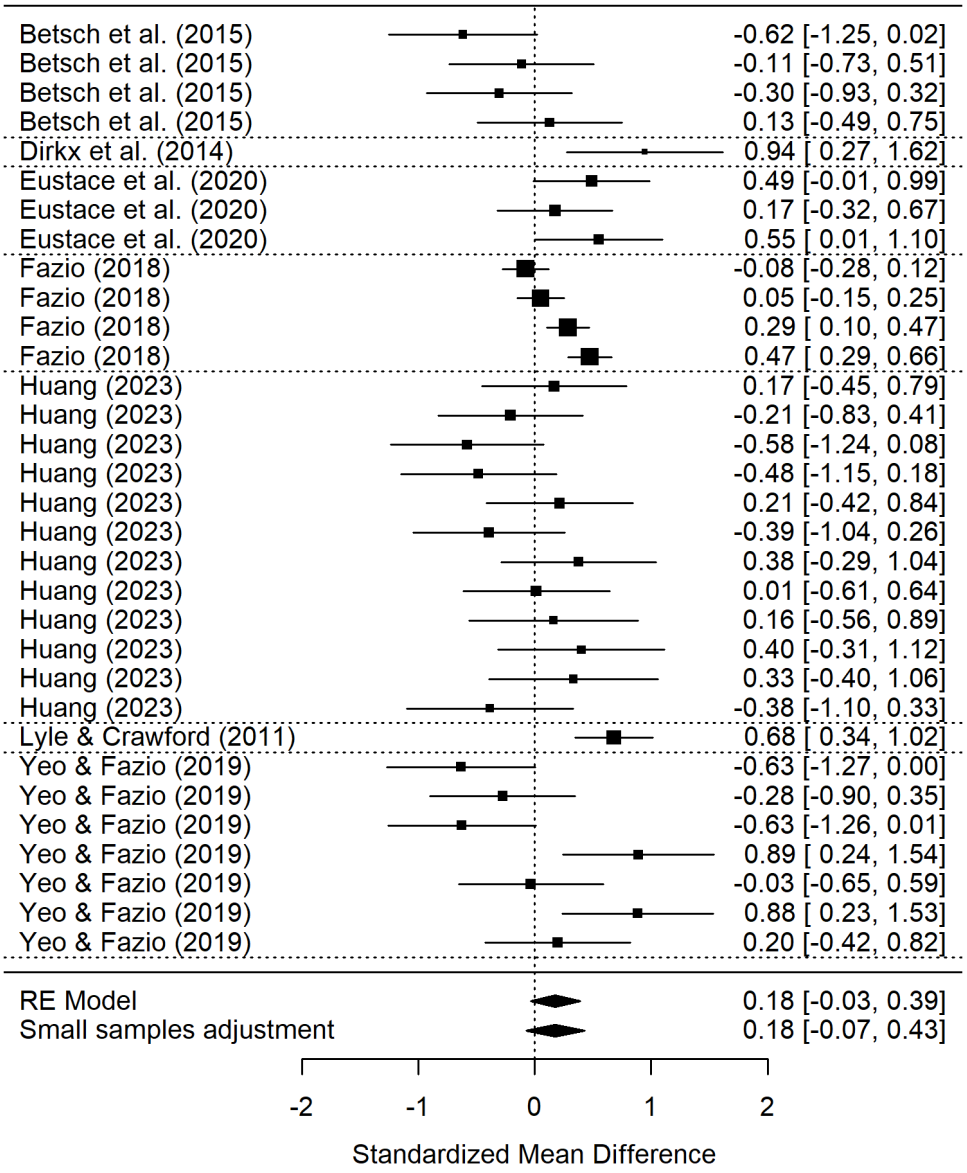
Where reported, participants were provided with three or four exposures to each item to be learnt, for example, in Fazio (2018) (Experiment 1) participants had either four study opportunities (SSSS) or one study and three test opportunities (STTT). Each study or testing session ranged from four (Yeo & Fazio, 2019) to eight minutes (Dirkx et al., 2014). For the

retrieval interval, 42% of the observations had a retrieval interval of less than one day, while the remainder were longer. Eustace et al. (2020), Betsch et al. (2015) and Lyle and Crawford (2011) provided feedback, and Yeo and Fazio (2019) gave feedback in one condition in their second experiment, while the other studies did not. Just under half of the studies (46%) received their retrieval opportunity immediately, while the remainder had it after a delay. Two studies were lab experiments (Fazio, 2018; Yeo & Fazio, 2019) while the remainder were performed in a classroom setting. Due to the nature of the testing effect most studies did not include an initial test, as even being tested before learning material can affect subsequent learning, this is known as the *pretesting effect* (Pan & Carpenter, 2023), the exceptions were Fazio (2018) and Dirkx et al. (2014). Fazio (2018) was also the only study that adopted a within-subject design.

Overall Effect size. Calculating the impact of testing versus restudying for mathematics learning, the overall weighted mean effect size is $g = 0.184$ (se = 0.095) (95% confidence interval [-0.069, 0.436]) (see Figure 2-8). Note that the 95% confidence intervals crossed the zero line. The hierarchical structure of the meta-analyses can allow us to see how much variance was associated with each level of the hierarchy. Firstly, I^2 = 30% of the variance was associated with the first level (sampling error), I^2 = 50% was associated with the second level (within study heterogeneity) (95% confidence interval [26.417, 83.001]) and I^2 = 20.105% associated with the third level (between-study heterogeneity) (95% confidence interval [0, 89.458]).

Figure 2-8

Retrieval versus Restudying - Forest Plot



Note. A forest plot displaying the weighted mean effect sizes for each effect in the subset. The size of the square is proportional to the sample size; the error bars represent the 95% confidence

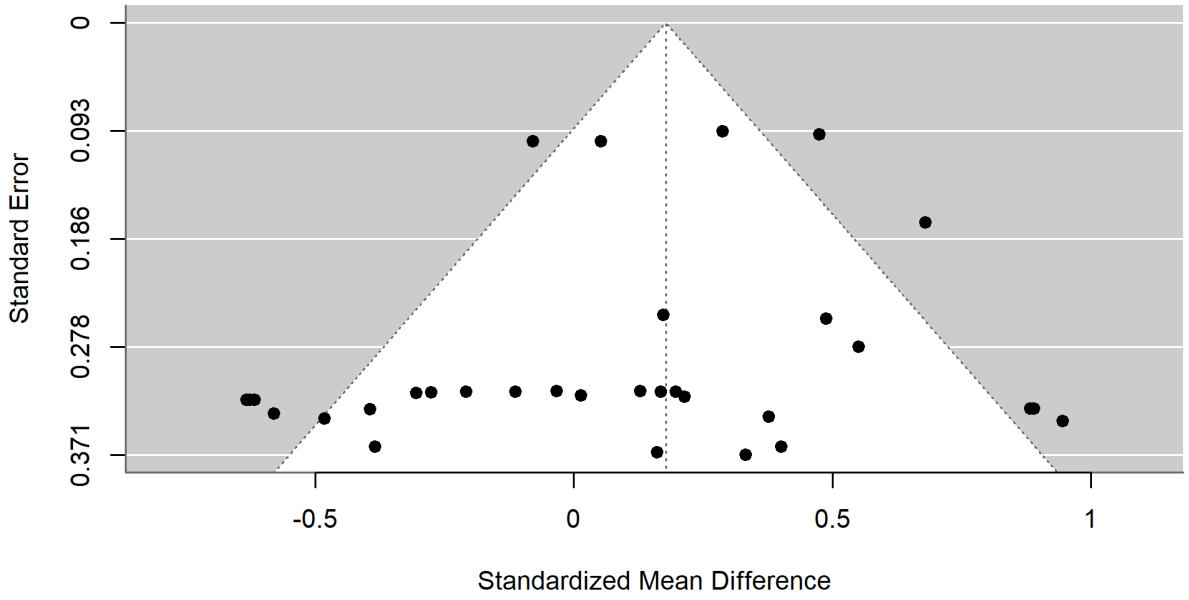
interval. The diamond at the bottom represents the overall weighted mean effect size before and after the small sample adjustment.

When the Cook's distances for each effect were calculated none were greater than 0.5 which suggested that no single study was likely to be highly influential. However, the difference in fit analysis shows that the removal of one study (Betsch et al., 2015) shifts the mean effect size by almost half a standard deviation ($DFFITS =$ -0.483), which changes the mean effect size to g = 0.239 (se = 0.096) (95% confidence interval [-0.038, 0.516]). These results suggest the lack of robustness is not due to any one studies inclusion as the 95% confidence interval always crosses zero regardless of which study is removed.

Risk of Bias. An insignificant Egger's regression test for funnel plot asymmetry (t = -1.279, d.f. = 30, p = 0.211) does not suggest publication bias resulting from insignificant results remaining unpublished. Additionally, there is not clear asymmetry in funnel plot (see Figure 2-9).

Figure 2-9

Retrieval versus Restudying - Funnel Plot



Note. A funnel plot showing the effect size of retrieval practice versus restudy against the standard error for each effect in the sample. Large asymmetry would suggest publication bias.

Our analyses were hampered by the small sample of studies that explicitly compare testing to restudy in mathematics, and the overall effect was smaller than previously found and was not robust.

2.4 Discussion

This meta-analysis had three purposes. First, to synthesize the current evidence regarding the efficacy of spaced versus massed practice with regards to mathematics learning. We found a robust small to medium effect of spacing, versus massed practice, overall, and in both subsets. This spacing effect was larger for material taught in isolation and smaller for material embedded into a course. Second, to investigate the efficacy of retrieval practice versus restudy within the

domain of mathematics learning. We found a small to medium effect of retrieval versus restudy; however, this was not robust as the 95% confidence interval crossed zero. Third, we intended to examine the effect of spaced retrieval practice versus massed retrieval practice, however, this was not possible due to a lack of studies.

Our first aim was to answer the question: does the spacing of mathematics practice lead to higher retention than massed practice? Our preregistered analysis looked at the weighted mean effect size of all studies that included a spaced versus massed manipulation. We found a small to medium effect of spacing (g = 0.26), and the 95% confidence interval suggests this is robust. However, this effect size is smaller than typically found effects of spacing in other domains such as L2 Language learning (g = 0.80, with a delayed post-test, Kim & Webb, 2022), however it is similar to more general reviews of spaced versus massed practice (g = 0.46, Donovan & Radosevich, 1999). This suggests that spacing is effective for mathematics learning, but some aspect of the material or the way it is taught may lead to reduced efficacy in comparison to other domains.

Furthermore, two subsets of studies with distinct paradigms contributed to this overall effect. Isolated learning studies followed a traditional spacing paradigm where massed practice is compared to spaced practice (over multiple sessions) before being tested after a specified delay. In contrast, course-embedded studies were embedded in a course structure, each individual item following a spacing schedule, however a crucial difference is that material at the start of the course may have a much longer retrieval interval than material at the end of the course. Additionally, the material is interleaved with the other questions. Interleaving provides a bonus effect of increasing learners' ability to discriminate between question types, a skill unnecessary in isolated learning studies, but relevant in course-embedded studies. Furthermore, due to the

cumulative nature of mathematics material, the end of the course typically builds on prior material, so the prior material may receive more practice than reported. For example, in a course on calculus, basic rules for differentiation may be tested explicitly at the beginning, but then also used to answer more complex questions at the end, resulting in additional practice.

For isolated learning studies, our analysis consisted of nine studies with thirty-five effect sizes. For isolated learning, there was a significant small to medium effect of spacing (g = 0.427), which was much larger than the overall spacing effect, however, the lower limit of the 95% confidence interval is closer to zero than the overall spacing effect. There were two studies whose removal would cause a significant change in the overall mean effect, however neither caused the confidence interval to cross zero. The increased experimental control available in isolated learning studies provides a purer measure of the spacing effect and is more comparable to past spacing effect meta-analyses such as the one conducted by Donovan and Radosevich (1999).

There are hints of publication bias in the isolated learning studies. The funnel plot is asymmetrical, there are no negative, or close to zero, effects for studies with larger standard errors than 0.3, while there are seven studies that found a significant effect with a standard error larger than 0.3, which hints at potential publication bias. However, while the funnel plot asymmetry hints at a possible under-reporting of smaller effects, neither Egger's regression test nor the trim and fill method suggested significant statistical evidence of publication bias. This suggests an overall lack of strong evidence for publication bias for isolated learning studies.

Seventeen course-embedded studies with twenty-six effect sizes were analysed. We found a robust significant effect of spacing for course-embedded studies (g = 0.24). The difference in fit analysis showed that the removal of no single study would cause the 95% lower

limit to drop below zero, providing further confidence in this study effect. As these studies implemented spacing into a course, they also benefited from interleaving which we would have expected to have produced an additional beneficial effect and increased performance on the post-test.

One potential explanation for the smaller mean effect (though not significantly smaller) in course-embedded studies could be linked to the study-phase retrieval hypothesis. The study-phase retrieval hypothesis suggests a key mechanism underlying the spacing effect is the retrieval of the initial study-phase. In a course setting in classrooms or as homework, students likely have prior material they can access during practice. If students have access to the materials, they may not have to retrieve the original memory, instead relying on external guides. Two isolated learning studies make clear statements as to which materials participants had access to, for example, “While working on the practice sheets, each student had access to a summary sheet containing examples, including solutions, from the introductory lesson.” (Barzagar Nazari & Ebersbach, 2019, pg. 291) and “Throughout each practice session, students could see their written work for practice problems that had appeared previously in the session.” (Emeny et al., 2021 pg. 1085). However, even in these cases it is difficult to understand to what extent participants made use of these materials. It has been previously noted that distributing practice may benefit medium performers best (Barzagar Nazari & Ebersbach, 2019). For these students, they may feel equipped to attempt the problem before checking the available materials, using them for feedback rather than undercutting the retrieval process. It would be valuable for future experiments to both control for and directly manipulate what materials students have access to, and when, during spaced practice of mathematics.

Within a course structure both the nature of mathematics learning and the lack of experimental control may have led to a smaller effect. The cumulative nature of mathematics learning means that, even in the massed conditions, simpler material may be practiced after a delay when more complex information is introduced, introducing an element of spacing into the massed condition. Further, sometimes spacing has been observed to increase accuracy when learners provide judgments of learning (Emeny et al., 2021; Logan et al., 2012), however, it is often clear whether you have successfully completed a mathematics problem so this beneficial effect may be minimised in the spaced condition. On the other hand, there are multiple reasons to predict the spacing effect could have been stronger within subset two. First, as these studies were embedded into a course structure, the learners may have been more intrinsically motivated as their grade counted for something, unlike most of the experiments in subset one. Second, in the spaced condition, the material was often also interleaved with material that had been previously learned, which could have provided a bonus ability for students to discriminate between question types and could have improved performance (Foster et al., 2019). Overall, there is no clear reason why spacing may be less effective when embedded in a course.

The analysis of studies comparing retrieval practice versus restudy sought to help us answer the question: does retrieval practice of mathematics lead to better learning (than just restudy)? It consisted of six studies with twenty effect sizes. Our analysis suggested the testing effect in mathematics may provide some advantage ($g = 0.184$); however, the 95% confidence interval crosses zero, which suggests that the effect is not robust.

We had planned to investigate whether the combination of spacing and retrieval lead to higher retention of mathematics knowledge (versus massed retrieval)? However, there was insufficient information in most studies to decide whether the spacing study required retrieval or

not. Ultimately, we deviated from our pre-registered retrieval moderator (yes or no) to retrieval (certain or uncertain). If it was a timed online test or under test conditions, we said it was likely to require retrieval, otherwise we could not be sure. The moderator analyses for whether retrieval was certain or uncertain were not significant.

Every effort was made to follow current best practices in meta-analysis (Steel et al., 2021) including pre-registering, checking for publication bias, quality indicators and using difference of fits analyses. We calculated the overall effect using robust variance estimation with correlated hierarchical effects. This allowed us to account for multiple comparisons within the same study (Pustejovsky & Tipton, 2022). We made systematic efforts to uncover unpublished studies; however, no authors we contacted reported any unpublished studies.

Given the current available literature, we could not state whether there is or is not a robust effect of testing for mathematics material. Similarly, the use of practice questions in mathematics and the lack of clarity as to whether learners were able to view past work meant that we could not accurately code whether studies required retrieval or not. We did, however, find significant evidence for a beneficial effect of spacing for mathematics material. This effect was smaller in studies where the material was embedded in a course, in comparison to stand alone experiments. Overall, the current state of the literature allowed us to answer one of our three questions, spacing is effective for mathematics material. However, we were unable to answer whether retrieval practice is effective for mathematics learning and how the testing and spacing effects interact. Future studies should aim to control, or directly vary, whether participants must retrieve the information during practice or have it available to them. We believed that while not differing statistically, the isolated and course-embedded subsets implement the spacing effect in two theoretically distinct ways. By carefully manipulating the order of practice problems, into

blocked, interleaved, and remote interleaved sets, Foster et al. (2019) were able to calculate the relative contributions to learning from spacing and interleaving. However, in their experiment the practice took place in a single session with no significant temporal delays between sets, so is not directly applicable to the difference between isolated and course-embedded material. Future experiments involving course-embedded spacing should take care to structure problem questions to allow the relative contributions of spacing and interleaving to be extracted. More testing effect studies that use testing effect best practices, rather than pushing the effect to find boundary conditions, are required to boost confidence in the utility of the testing effect to benefit mathematics material.

2.5 Conclusion

We found robust evidence for a positive small to medium effect of spaced rather than massed practice for mathematics learning. This effect was numerically larger for studies where participants learnt isolated material in contrast to when the material was embedded in a course, but importantly, the effect was robust in both subsets. We found a small to medium mean effect of testing versus restudy; however, this was not robust. Overall, there is sufficient evidence to promote the use of spacing for mathematics learning to improve mathematics learning. Given the robust nature of retrieval practice in other domains, there is insufficient evidence to suggest that retrieval practice is not a useful pedagogical tool for mathematics learning, however, more evidence is required to ensure the robustness of the testing effect.

2.6 Appendix

PICOS table

Table 2-*2*

PICOS Table

PICOS Term	Inclusion	Exclusion
Population	Any age range or education level.	Clinical populations (classroom studies where particular participants have additional needs will be included, but not exclusively clinical samples)
	Lab or classroom based.	
Intervention	An intervention that bases its design on the use of the spacing effect, retrieval effect, or a combination of the two and uses mathematics material.	Studies which use multiple subjects (for example Science questions and mathematics questions) where the results of the mathematics section is not reported in sufficient detail to calculate an effect size.
Comparators	Massed versus spaced practice	No comparator or control condition
	Greater versus lesser spacing	
	Retrieval versus restudying	

PICOS Term	Inclusion	Exclusion
Outcome	Performance on a post-test	
Study design	Includes a post-test after the condition or control intervention	Non-experimental designs No comparator or control condition
	Is the data already reported in another paper included in the review?	

Note. Description of inclusion and exclusion criteria following the Population, Intervention, Comparison, Outcomes and Study (PICOS) framework

3 More Steps, Same Effect: Spacing Increases the Retention of Mathematics Procedures of Varying Complexity

Additional information for journal style thesis:

The following chapter has been submitted at a journal; however, it was not accepted initially and will be rewritten with the hopes of submission.

Statement of authorship:

This paper was authored by Ewan Murray and reviewed, edited, and supervised by Aidan J. Horner and Silke M. Göbel. Ewan Murray performed all data analysis, and the results were reviewed by Aidan J. Horner and Silke M. Göbel.

Data availability:

The data and code required to reproduce this manuscript, and all analyses is available on the OSF (***https://osf.io/cex8g/***).

Abstract

Spacing, distributing practice over time rather than in a single session, often benefits long-term memory of simple material. However, it is less clear whether spacing is effective for more complex material. As more educators harness the spacing effect, it is important to know under what conditions it is most effective. We investigated the impact of procedural complexity on the efficacy of spacing, by varying the number of steps in arithmetic procedures. Participants were taught two procedures, either in a single session (massed) or over three sessions spanning three consecutive days (spaced). Experiment one compared learning a two-step with a three-step procedure. Spaced practice led to significantly higher performance, relative to massed practice, with no evidence for a difference in the spacing effect as a function of procedural complexity. However, there was also no evidence for a difference in performance between the two procedures, suggesting the three-step procedure was not sufficiently more complex than the two-step procedure. Experiment two compared learning a two-step with a five-step procedure. We again saw a significant spacing effect, as well a main effect of complexity, with performance in the five-step procedure being significantly lower than the two-step procedure. As in Experiment one, we found no evidence for an interaction between procedural complexity and the spacing effect. Our results show that spacing benefits the learning of arithmetic procedures. Critically, we also show that the spacing effect is not negatively impacted by the procedural complexity of the procedure learnt.

Keywords: Mathematics learning, Memory, Spacing effect, Distributed practice, Spaced Retrieval Practice, Complexity, Procedure

More Steps, Same Effect: Spacing Increases the Retention of Mathematics Procedures of Varying Complexity

When learning mathematics, a student will often have to recall and use material from the past. This material could be from the previous lesson, semester, or year. If, during a lesson, students are unable to recall these procedures or concepts, they must take time away from learning new material to relearn them. One potential solution to this problem is to distribute practice over time and over multiple sessions, as opposed to massed practice in one session. The change in retention due to the distribution of practice over time is termed *the spacing effect*. While the spacing effect often benefits long-term memory of simple material (Cepeda et al., 2006), it is less clear that spacing is effective for more complex material (Donovan & Radosevich, 1999). We chose to use mathematics material to investigate this relationship between complexity and spacing as we could manipulate the number of steps in a procedure and test prior knowledge for the skills required in each step. In this study we ran two experiments to investigate how procedural complexity, defined by the number of steps in a procedure, affects the efficacy of spaced retrieval practice versus massed retrieval practice.

The spacing effect is a robust phenomenon in learning and memory research. Previous meta-analyses have found large beneficial effects to learning outcomes across hundreds of studies (Cepeda et al., 2006; Donovan & Radosevich, 1999; Latimier et al., 2021), and it has been subject to multiple reviews (Delaney et al., 2010; Küpper-Tetzel, 2014; Maddox, 2016). In the domain of mathematics, there have been mixed results. Significant benefits have been reported for learning arithmetic facts (Schutte et al., 2015), algebra (Gay, 1973), geometry (Yazdani & Zebrowski, 2006), simple procedures (Rohrer & Taylor, 2006), and more complex bodies of knowledge such as calculus (Hopkins et al., 2016; Lyle et al., 2020, 2022). Significant

positive effects have been found across age groups including primary school (Chen & Kalyuga, 2019), secondary school (Emeny et al., 2021; Barzagar Nazari & Ebersbach, 2019) and higher education (Hopkins et al., 2016; Lyle et al., 2020, 2022).

Nevertheless, there have also been multiple studies that did not find a significant positive effect, and sometimes a negative effect, of spacing for mathematics learning (Beagley & Capaldi, 2020; Ebersbach & Barzagar Nazari, 2020a; Rohrer & Taylor, 2006). One reason may be participants' performance during practice. During exploratory analyses, Barzagar Nazari and Ebersbach (2019) found that distributed practice was most effective for students of medium performance, while high performers may already be at ceiling and lower performers may never achieve sufficient proficiency in the task to benefit from spacing. Alternatively, the results in Barzagar Nazari and Ebersbach (2019) could be explained by whether participants had to retrieve the information or not. During the task participants had access to a summary sheet containing solved examples. High performers could have been at ceiling, but medium performers may have experienced sufficient success to practice without relying on the summary sheet. In contrast, low performers may have relied heavily on the summary sheet without having to retrieve the information. Higher attrition in the spaced condition, in comparison to the massed condition, has also been an issue, leading analyses to be dropped (Barzagar Nazari & Ebersbach, 2018). In the methods section we describe how we have attempted to avoid issues with retrieval, performance during practice and attrition bias in the two experiments. In summary, there is evidence that spaced practice can be beneficial for mathematics learning, however, there may be additional factors that limit the effects of spacing or require the practice to be implemented differently. We will focus on one such factor in the following experiments, the complexity of the material.

In their meta-analysis looking at spacing across a variety of tasks, Donovan and Radosevich (1999) found a significant negative correlation between the *overall complexity* of a task and the efficacy of spacing. They defined overall complexity as "the degree to which the task requires a number of distinct behaviours, the number of choices involved in the performance of the task, and the degree of uncertainty involved in performance of the task" (p. 798). In our study we defined complexity procedurally, by counting the number of steps required to solve a problem. This has been used in previous mathematics learning experiments that manipulated procedural complexity (Mattis, 2015; Vincent & Stacey, 2008). We reasoned that the number of steps would map onto the number of distinct behaviours aspect of Donovan and Radosevich (1999)'s definition of overall complexity and hypothesized it might affect the efficacy of spacing for procedures with more steps. We consider two theories of the spacing effect and how they may interact with complexity.

Firstly, the *study-phase re*trieval hypothesis (Thios & D'Agostino, 1976) suggests that the benefit of spacing arises from the retrieval of the initial study-phase, with each successful retrieval creating additional routes for retrieval, or multiple copies of the memory, leading to better future recall. If more complex information is either more difficult to retrieve or contains multiple parts that can be independently forgotten, then spacing may be less effective for more complex material, because participants will retrieve the original study phase less frequently or incorrectly. The effort required to successfully retrieve material also affects the outcome of spacing, where a successful retrieval that is more effortful would lead to a larger effect (Pyc & Rawson, 2009).

Secondly, the *deficient processing account* suggests that during massed learning the material is more shallowly processed due to seeing the stimuli frequently over a brief period of

time (Hintzman, 1974). In contrast, in the spaced condition, the amount of processing resets at the start of each new session, which leads to higher average processing over multiple trials and in turn, greater quality learning overall. The complexity of the material may interact with the amount of processing, in turn affecting any deficient processing in the massed complex material condition. While a simple item may quickly be less processed when seen frequently, perhaps a more complex item will require more intentional processing and will not be affected by deficient processing. If that were the case, there would be no additional benefit of spacing due to deficient practice for more complex items.

We investigated the impact of procedural complexity on the efficacy of spacing using a two-by-two experimental design. We varied the number of steps in artificial arithmetic procedures (two versus three in experiment one, two versus five in experiment two) and the practice schedule (one massed versus three distributed sessions). This had several benefits. The artificial nature of the procedures minimised the impact of prior knowledge, as participants will not have seen these procedures before. We were able to check that participants would be proficient in all the basic arithmetic skills ensuring that each could be counted as a single element. We ensured that it is not possible to skip steps, which has been an issue in prior attempts to define complexity (Mattis, 2015). Increasing the number of steps exponentially increases the number of ways to misremember the order of the steps. While we lose ecological validity, we believe the additional control gained over prior knowledge make this a beneficial choice.

We had three hypotheses. Firstly, spaced retrieval would lead to a greater retention than massed retrieval across all conditions. We did not expect the spacing effect to fully disappear or negatively affect learning. This hypothesis predicts a significant main effect of spacing.

Secondly, participants would have lower retention of higher complexity material. We predicted that as there were more connected steps to recall and use, participants in the higher complexity condition would have lower performance on the post-test. This hypothesis predicts a significant main effect of complexity. Thirdly, there would be an interaction between spacing and complexity. Specifically, there would be a significantly smaller effect of spacing in the high complexity condition than in the low complexity condition.

We also measured participants' working memory, arithmetic fluency, retrieval accuracy during practice, and mathematics anxiety to check that the initial groups were equal in these variables to ensure this did not affect the outcome.

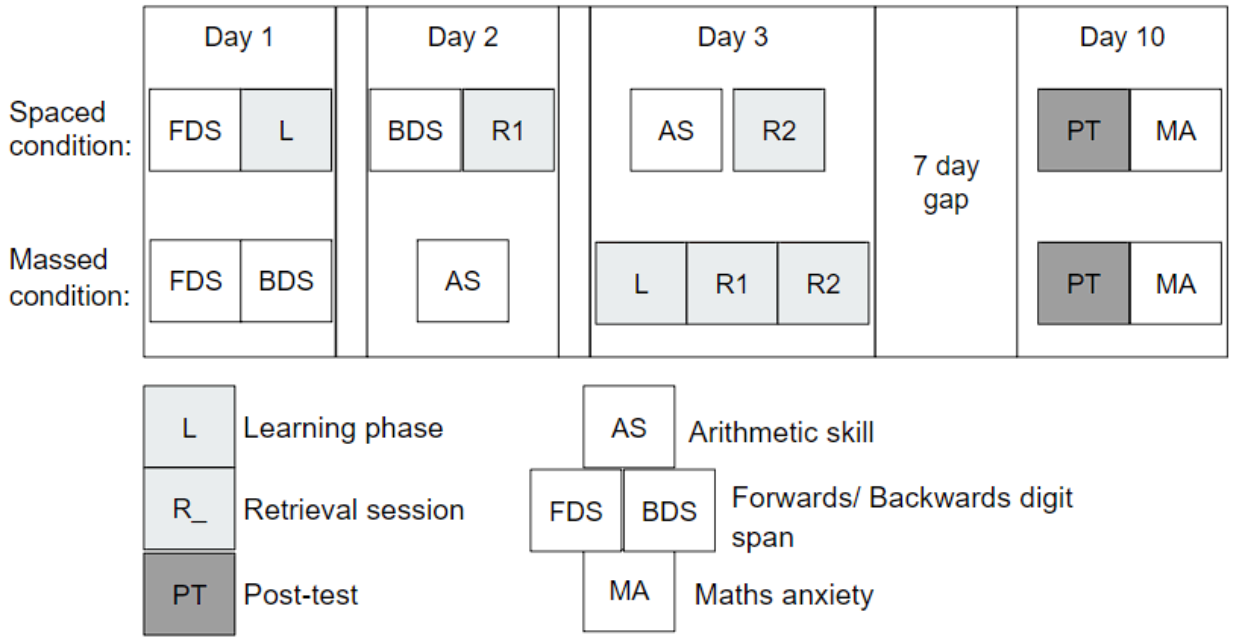
3.1 Experiment One

3.1.1 Method

The study was implemented as a mixed factorial design, two factors each with two levels. Within participants, procedural complexity was manipulated; participants learnt both a two-step and three-step procedure. Between participants, the schedule of learning was manipulated; participants either learnt the procedures in one longer session (massed) or over three sessions split over three days (spaced). Each participant completed eight tasks over four sessions. This included a forwards and backwards digit span task, arithmetic fluency, one procedure learning session, two retrieval sessions, a post-test, and a mathematics anxiety questionnaire. The order in which participants completed the tasks depended on their scheduling condition (see Figure 3-1).

Figure 3-1

Experimental Design



Note. A diagram outlining the overall procedure of the experiment.

3.1.2 Participants

Using the R module *SuperPower* (Lakens & Caldwell, 2021) to simulate a two-by-two Mixed ANOVA suggested that 68 participants (34 per condition) would be sufficient to detect an $\eta^2 = 0.11$ main effect of spacing, which is similar to results found in prior experiments. Bego et al. (2017) found a significant main effect of spacing $\eta^2 = .147$ and similarly Hopkins et al. (2016) found a main effect of $\eta^2 = 0.099$. This was, however, only sufficient power to detect the main effects of complexity and spacing. As this is a novel paradigm, and there is little prior evidence for the effect size of the interaction, we only initially powered the main effects, due to time and resource constraints.

All participants provided their informed consent and both experiments were approved by the University of York's Ethics board.

One hundred and one participants began the experiment; however, a total of thirty-one participants dropped out before completing the fourth session (30.69% attrition rate). Twelve participants dropped out of the experiment before being assigned to either the massed or spaced routine, we predicted that when asked to sign up to a time to do the experiment participants would realise, they did not have sufficient time to complete it and drop out. The remaining attrition was evenly distributed between the massed and spaced condition. Nine participants began the massed condition before dropping out and ten participants began the spaced condition before dropping out, suggesting little attrition bias. The final sample consisted of seventy undergraduate students from a British university (three reported to already have a bachelor's degree, however, based on their age they may have misread the question and thought they needed to put what course they were doing). The final sample was heavily biased towards female participants (87%), with the remainder identifying as male, participants were aged between 18 and 21 years old and the mean age was 19.11 years (SD = 0.79 years).

3.1.3 Material

Learning arithmetic procedures. Participants were taught two arithmetic procedures, where each step required participants to perform an operation using a particular operand (i.e., in "Multiply by two" the operator is "Multiply" and the operand is "two") (see Table 3-1). The procedures consisted of only addition, subtraction, multiplication, division and squaring a number as the operations. In the first experiment, one procedure had two steps, and the other had three steps. The initial step was always "Add the two numbers together" or "Multiply the two numbers together." This ensured the order of the two numbers did not matter and the next step only had to deal with one input.

Table 3-1

Procedures Learnt by Participants

Counter balance	Two-step	Three-step*	Five-step**
One	Add the two numbers together	Add the two numbers together	Add the two numbers together
	Divide your answer by three	Divide your answer by three	Divide your answer by three
		Square your answer	Subtract two from your answer
			Square your answer

Counter balance	Two-step	Three-step*	Five-step**
Two	Multiply the two numbers together	Multiply the two numbers together	Multiply your answer by ten
	Subtract ten from your answer	Subtract ten from your answer	Multiply the two numbers together
		Divide by two	Subtract six from your answer
			Divide your answer by four
			Square your answer
			Add one to your answer

Note. Table displaying the procedures used in the experiments. * = used only in experiment one, ** = used only in experiment two

In both the learning session, and the retrieval sessions, participants were cued with the instructions “Apply the n-step procedure to the numbers a and b” with n being either “two” or “three” and *a* and *b* two integers such that the result of applying the procedure is a whole number. In the learning session, the steps were displayed while participants applied the procedure, while in the retrieval sessions the participants had to retrieve the steps to apply the procedure. In all three sessions, when an incorrect answer was submitted participants were given

feedback in the form of the correct answer and were shown the steps of the procedure again. Each session consisted of nine trials for the two-step and nine trials for the three-step procedure. Retrieval accuracy was defined as the mean accuracy across the two retrieval sessions.

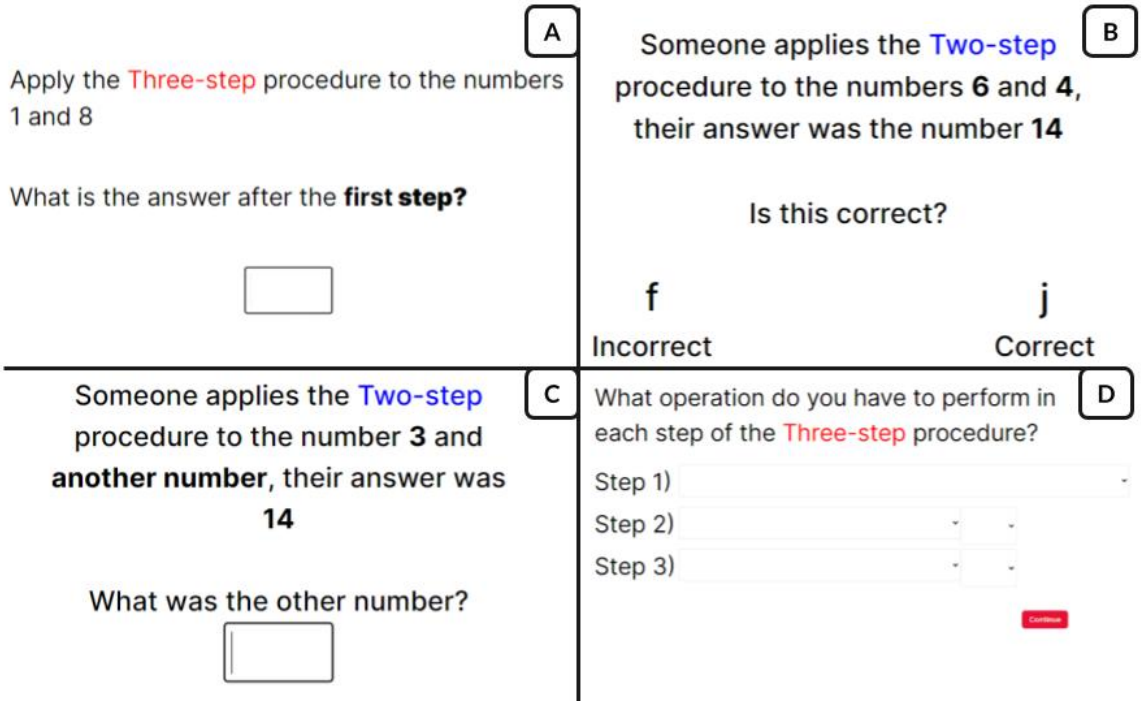
Two attention checks were presented during each learning, retrieval, and post-test task. The attention checks were visually similar to the main task; however, participants simply had to type in the number presented. We preregistered that if participants failed two attention checks during any one session, they would be removed, however, no participant failed both checks in experiment one.

The critical post-test contained four sub-tasks. The participants went through the tasks for one procedure, then the next, with the order counterbalanced across participants. We used four tasks to capture a more rounded view of participants memory for the procedures. The first task consisted of six questions identical to the retrieval task, referred to as *basic trials,* but with no feedback, where participants either performed the procedure in full, or gave the answer after an intermediate step (see Figure 3-2). In the second task participants were presented with the procedure, the inputs, the answer, and were then asked if it was correct. This task is later referred to as *correct (y/n)* trials (see Figure 3-2 -B). In the third task, *reverse procedure trials,* participants were given the procedure, one of the two inputs and the answer, and were then asked to reverse the procedure to find the other input (see Figure 3-2 - C). Finally, participants were asked to choose the correct steps from drop-down boxes (see Figure 3-2 - D). For the *recognise steps task*, participants could get one point for the first step and two points for each subsequent step, one for the operand and one for the operator. For each procedure, the final score on the recognise steps task is scaled, such that remembering all the steps is worth three points. The final

score for the post-test was calculated as the sum of the raw scores in the first three sub-tasks plus the scaled score for the drop–down procedure task, for a maximum potential score of fifteen.

Figure 3-2

Post-test Sub-tasks



Note. Four screenshots from the experiment outlining the four sub-tasks in the post-test. First Basic trials where participants simply apply the procedure in full or after a certain number of steps(A). Correct (y/n) trials where participants are given the procedure, inputs and answer and must see if it is correct (B). Reverse Procedure trials where participants are given the answer and one of the inputs and tasked to work out the other input (C). Recognise steps trials where participants are asked to select the correct steps from a drop-down box (D).

Individual difference measures. Working memory was measured by forwards and backwards digit span tasks. The two tasks were based on the Wechsler memory scale - third edition (WMS-III) (Wechsler, 1997). During the task, participants heard a spoken series of numbers. Beginning at two digits, participants either had to recall them forwards or backwards and type them into a box on the screen. Each round had two chances to succeed at repeating the numbers correctly, if they succeeded (in at least one of the two trials) then they would move onto the next round where an additional number was added. If not, then the task would end, and their final score was calculated as the total number of trials they answered correctly during the task. The maximum possible number of digits presented in one trial (span) was nine for the forwards digit span and eight for the backwards digit span. The maximum score was 18 and 16 for the forwards and backwards digit span, respectively.

Arithmetic fluency was measured using a version of the Math4Speed task (Loenneker et al., 2022) which was altered to be usable online. It consisted of fifty addition, fifty subtraction, fifty multiplication and fifty division questions. Participants had two minutes for each arithmetic type to complete as many questions as possible. The sum correct answers across each operation were used as our measure of arithmetic fluency.

Mathematics anxiety was measured using an online version of the Mathematics Anxiety Scale–UK (MAS-UK) (Hunt et al., 2011). This scale was specifically designed for the British undergraduate student population, which matched our target population. It consisted of twenty-three hypothetical scenarios designed to measure participants' mathematics anxiety. For each scenario, participants rated their predicted anxiety on a on a five-point scale (with the answer options: 'not at all', 'slightly', 'a fair amount', 'much', 'very much'), where a higher score meant they felt more anxious at the prospect of doing that activity. For example, participants were

asked to imagine how anxious they would feel when a situation occurs such as “Having someone watch you multiply 12 x 23 on paper”. The final mathematics anxiety score was the sum of their answers.

3.1.4 Procedure

We recruited participants through the University of York’s human participant pool via Sona Systems (***https://www.sona-systems.com/)***. The experiment was created and hosted online through the platform Gorilla (Anwyl-Irvine et al., 2020) and upon completion of the study the participants were granted course credit. Once directed to the experiment, participants were presented with a consent form. Once they consented, they answered a series of demographic questions and were directed to choose times when they would complete the study. This was done so that anyone who misread the instructions and would not be able to complete the study would be removed here. They were then randomly assigned to either a massed or spaced schedule to complete the remaining tasks, using Gorilla’s built in randomizer.

The first three sessions were scheduled over three consecutive days then a final session seven days after the third session. In the massed condition participants completed the learning phase and then the two retrieval phases on day three; and then the post-test on day ten. In the spaced condition participants completed the learning phase on day one, the first retrieval phase on day two, the second retrieval phase on day three, and then the post-test on day ten. This ensured the retrieval interval for both the massed and spaced conditions was equivalent.

In another study involving spacing, there was more attrition in the spaced condition, than the massed condition, which meant the original analyses could not be run (Barzagar Nazari & Ebersbach, 2018). To reduce the possibility of this for the present study we ensured that participants in the massed and spaced condition both had to complete four sessions over ten days.

The digit span and arithmetic fluency tasks were used to pad the remaining sessions to ensure that participants had to return to the experiment an equal number of times across both conditions.

3.1.5 Analysis

We performed a two (massed versus spaced) by two (two-step versus three-step) mixed ANOVA on accuracy on the post-test. This was pre-registered (***https://osf.io/td5h8***). As the sample size is greater than fifty participants Q–Q plots were used to assess the normality of the data. If the points fall approximately along the line of the Q–Q plot, then normality was assumed. Levene's test for homogeneity of variance was used to test for heteroscedasticity. Box's M-test was used to check for Homogeneity of Covariance. Any assumptions that were broken were reported. For completeness, post-hoc t-tests with Bonferroni corrections were run to investigate where the difference arose. Each ANOVA was also run using the *BayesFactor* R package (Morey & Rouder, 2023) with the default Jefferies priors to calculate a Bayes Factor for each effect. We used an alpha level of .05 for all statistical tests, except for Box's M test which used a value of .001, due to its sensitivity.

Participants' data were excluded if they did not complete all the sessions or if they failed both attention checks in the learning, retrieval, or post-test phases of the experiment. Outliers more than three times the interquartile range were removed from the main analysis. Outliers more than one and a half times the interquartile range were inspected and the decision to include or exclude them explained.

As exploratory analyses we ran multiple t-tests to see if working memory, mathematics anxiety, arithmetic fluency, or accuracy during the retrieval sessions significantly differed between those in the massed or spaced conditions. The p-values for these analyses were adjusted using the Bonferroni correction for multiple comparisons.

3.1.6 Results

A two (spaced versus massed) by two (two-step versus three-step procedure) mixed ANOVA was performed, with percentage overall score on the post-test as the dependent variable (see Table 3-2). We ran the appropriate tests, listed in the methods, to ensure the data met the assumptions for the ANOVA. These conditions were met, with the exception of potential outliers more than 1.5 times the interquartile range from the first or third quartile, there were no extreme outliers. These outlying participants (n = 6, all in the spaced condition) performed worse than others, however, they did not fail attention checks and appeared to complete the rest of the study properly. Therefore, they have not been removed. In further exploratory analyses (not reported), removing the outliers appears to increase the effect of spacing, but did not affect the other main effect or interaction term.

Table 3-2

Experiment One Descriptive Statistics

Spacing	Procedure	Overall Score	Retrieval Accuracy	Working Memory	Mathematics Anxiety	Arithmetic Fluency
Massed	Two-step	0.561 (0.276)	0.843 (0.109)	19.971 (3.362)	51.118 (10.117)	69.265 (23.839)
	Three-step	0.608 (0.318)	0.863 (0.120)			
Spaced	Two-step	0.698 (0.229)	0.813 (0.079)	18.972 (3.468)	52.361 (14.333)	70.389 (23.626)

Spacing	Procedure	Overall Score	Retrieval Accuracy	Working Memory	Mathematics Anxiety	Arithmetic Fluency
	Three-step	0.734 (0.267)	0.844 (0.086)			

Note. Table displaying the mean and standard deviation for the percentage score on the post-test, Retrieval Accuracy (the mean percentage accuracy across both practice retrieval sessions) and the scores on the working memory, mathematics anxiety, and arithmetic fluency tasks

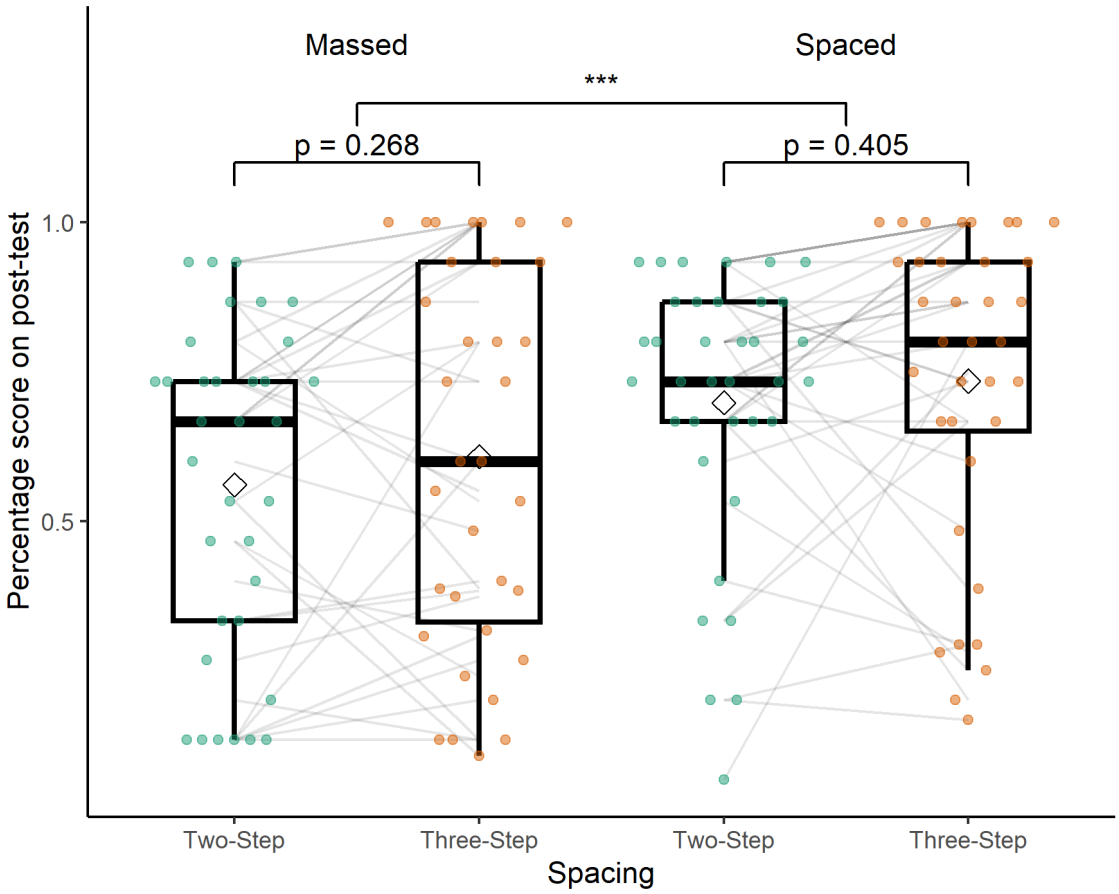
A two (spaced/massed) by two (two-step/ three-step procedure) mixed ANOVA was performed, with percentage accuracy on the post-test as the dependent variable (see Figure 3-3). There was a significant main effect of spacing, $F(1,68) = 5.135$, $p = 0.027$, $\eta^2 = 0.056$ ($BF_{10} = 2.291$). The percentage overall score on the post-test was significantly higher in the spaced ($M = 0.716$, $SD = 0.248$) than in the massed condition ($M = 0.585$, $SD = 0.297$). There was no significant main effect of complexity, $F(1,68) = 1.926$, $p = 0.17$, $\eta^2 = 0.006$ ($BF_{10} = 0.428$). There was no evidence of an interaction between spacing and complexity, $F(1,68) = 0.035$, $p = 0.853$, $\eta^2 = 0$ ($BF_{10} = 0.251$). We therefore saw a clear effect of spacing, however the lack of difference between the two-step and three-step procedure suggests that we may not have systematically varied complexity sufficiently.

Finally, we assessed whether the spacing effect was present in both the complexity conditions separately. Post-hoc t-tests revealed a significant difference between massed ($M = 0.561$, $SD = 0.276$) and spaced ($M = 0.698$, $SD = 0.229$) performance on the post-test for the two-step procedures ($t(34) = -2.258$, $p = 0.027$, $BF_{10} = 2.146$) after Bonferroni adjustment.

However, there was no significant difference between massed and spaced three-step procedures ($t(34) = -1.79$, $p = 0.078$, $BF_{10} = 0.969$), after Bonferroni adjustment.

Figure 3-3

Experiment One - Main effect of Spacing



Note. A boxplot showing the percentage score on the post-test by spacing condition and number of steps in the procedure. The significance stars represent: * p < .05, ** p < .01, *** p < .001

Table 3-3

Experiment One Break

down of Post-test Subtasks

Spacing	Procedure	Basic	Correct (y/n)	Recognise Steps	Reverse Procedure
Massed	Two-step	0.515 (0.361)	0.755 (0.310)	0.520 (0.428)	0.500 (0.500)
	Three-step	0.549 (0.401)	0.745 (0.308)	0.659 (0.427)	0.539 (0.539)
Spaced	Two-step	0.727 (0.314)	0.880 (0.213)	0.528 (0.460)	0.630 (0.630)
	Three-step	0.662 (0.337)	0.778 (0.276)	0.885 (0.243)	0.685 (0.685)

Note. Table displaying the mean and standard deviation for the percentage score on the post-test, broken down by the sub-tasks.

3.1.7 Exploratory Analyses

We ran four t-tests, with Bonferroni correction, to test if there was a significant difference of working memory, arithmetic fluency, mathematics anxiety, or accuracy across the two retrieval tasks between the massed and spaced conditions. We found no significant difference between any of the variables (see Table 3-4).

Table 3-4

Experiment one exploratory variables t-test

	Statistic	d.f.	p	p.adj
Arithmetic Fluency	-0.20	67.70	0.84	1.00
Mathematics Anxiety	-0.42	63.06	0.68	1.00
Retrieval Accuracy	1.31	59.85	0.20	0.78
Working Memory	1.22	67.95	0.23	0.90

Note. Table displaying the results of t-tests to investigate if there are any significant differences between massed and spaced groups for experiment one. The p values are adjusted using the Bonferroni correction.

3.1.8 Discussion

In this experiment we tested how changes to the number of steps in a procedure affected the efficacy of spaced retrieval practice. Our first hypothesis was that participants would have lower retention of high complexity material. However, the main effect of complexity was not significant and, numerically, participants appeared to perform better in the higher complexity condition. Secondly, we predicted that across all levels of complexity spaced retrieval would lead to a greater retention than massed retrieval. This was not the case. There was an overall significant effect of spacing, although, post-hoc results only showed significant evidence for a spacing effect in the lower complexity condition and the evidence was not significant in the higher complexity condition. The Bayes factor for the comparison between spaced and massed three-step procedures was close to one. This evidence suggests that it is equally likely that there

is a difference as that there is not a difference. Finally, based on previous evidence we predicted there would be a larger spacing effect for lower complexity material than higher complexity material. We found no evidence for an interaction between spacing and complexity. In the next experiment we aimed to increase the difference in complexity by adding additional steps to the higher complexity procedure.

3.2 Experiment Two

3.2.1 Method

3.2.2 Participants

Initially, one hundred undergraduate students from a British university were recruited online, through Sona. Of those, sixty-eight completed the study (32% attrition rate). Five participants dropped out before being assigned a condition. Otherwise, the attrition rates were similar across conditions, sixteen massed and eleven spaced participants failed to complete the study. No participant failed both attention checks in experiment two. Three participants in the spaced condition were extreme outliers and were removed. Of the remaining sixty-five participants fifty-seven identified as female, five as male and three participants identified as neither male nor female. The participants were aged between 18 and 21 years old and the mean age was 19.09 years (SD = 0.91).

3.2.3 Material

Following the results of experiment one, experiment two was designed to increase the difference in procedural complexity by adding two additional steps to the higher complexity procedure (see Table 3-1). Otherwise, the materials were the same.

3.2.4 Procedure

Experiment two followed the exact same procedure and data analysis plan as experiment one, however, the three-step procedure was replaced with a five-step procedure (see Table 3-1). This was pre-registered (***https://osf.io/bpfqm***).

3.2.5 Results

A two (spaced versus massed) by two (two-step versus five-step procedure) mixed ANOVA was performed, with percentage overall score on the post-test as the dependent variable (see Table 3-5 and Figure 3-4). Three extreme outliers (three times the interquartile range from the first or third quartile) were removed and the assumptions for normality and homogeneity of covariances were met. Levene's test indicated that the variances were homogeneous across the five-step procedure groups, $F(1,63) = 0.82$, $p = .365$, but not across the two-step procedure groups $F(1,63) = 10.64$, $p = .002$. This was due to a ceiling effect in the spaced two-step condition. We continued with the planned analysis.

Table 3-5

Experiment two descriptive statistics

Spacing

Spacing	Procedure	Retrieval Accuracy	Overall Score	Working Memory	Mathematics Anxiety	Arithmetic Fluency
Massed	Two-step	0.949	0.659	19.088	54.412	62.853
		(0.106)	(0.225)	(3.297)	(12.225)	(18.868)

Spacing	Procedure	Retrieval Accuracy	Overall Score	Working Memory	Mathematics Anxiety	Arithmetic Fluency
	Five-step	0.959	0.495			
		(0.126)	(0.251)			
Spaced	Two-step	0.969	0.852			
		(0.102)	(0.108)	19.824	49.029	75.794
	Five-step	0.928	0.7	(3.407)	(14.379)	(24.733)
		(0.095)	(0.228)			

Note. Table displaying the mean and standard deviation for the percentage score on the post-test

We believe that the lack of homogeneity of variance has not affected the results of the mixed two-way ANOVA for two reasons. Firstly, in an analysis (not preregistered) available in the supplementary materials we ran a Monte-Carlo simulation to investigate the impact of the ceiling effect in the spaced two-step condition. As our variances were homogeneous in the first experiment and in three out of the four conditions in the second experiment, we assumed that if there was no ceiling the variance would be similar in the spaced two-step condition. We generated 10,000 data sets based on this assumption and 10,000 data sets where a ceiling was artificially imposed at $y = 1$. The results of imposing the ceiling effect made little difference to the results and they were similar to the case where homogeneity was guaranteed. Secondly, we ran a non-parametric version of the mixed ANOVA (not preregistered) using the WRS2 R package (Mair & Wilcox, 2020) (see Table 3-6), the results of this analysis mirror the results

provided by the original mixed ANOVA, both main effects were significant and there was no interaction. We continued with the original ANOVA as it was possible to calculate effect sizes and we believe the lack of homogeneity, in one condition, did not impact the analysis in a meaningful way.

Table 3-6

Experiment two Robust ANOVA

	df1	df2	Q	p
Spacing	1	36.209	27.469	< 0.001
Procedure	1	32.908	20.212	< 0.001
Interaction	1	32.908	1.093	0.303

Note. The results of a robust between-within subjects ANOVA on the trimmed means using the WRS2 R package

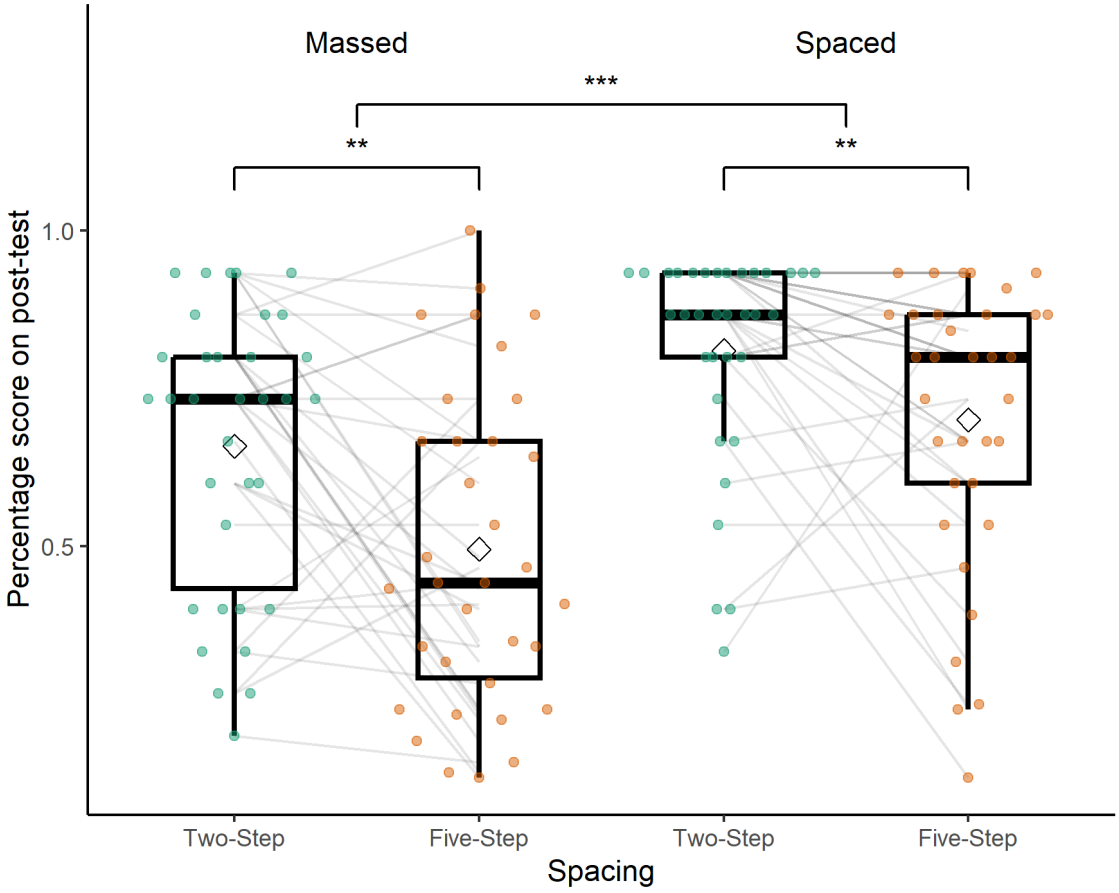
There was a large and significant main effect of spacing, $F(1,63) = 22.523$, $p < 0.001$, $\eta^2 = 0.184$ ($BF_{10} = 774.526$). The percentage overall score on the post-test was significantly higher in the spaced condition ($M = 0.776$, $SD = 0.193$) than in the massed condition ($M = 0.577$, $SD = 0.251$). There was a significant medium to large main effect of complexity, $F(1,63) = 24.308$, $p < 0.001$, $\eta^2 = 0.124$ ($BF_{10} = 3970.715$), such that participants performed better on the two-step procedures than the five-step. There was no evidence of an interaction between spacing and complexity, $F(1,63) = 0.036$, $p = 0.85$, $\eta^2 = 0$ ($BF_{10} = 0.281$).

Post-hoc t-test results on the spacing condition suggest that there was a significant difference between massed (M = 0.659, SD = 0.225) and spaced (M = 0.852, SD = 0.108) performance on the post-test for the two-step procedures ($t(34) = -4.462$, $p < 0.001$, BF_{10} = 382.014) after Bonferroni adjustment and a significant difference between massed (M = 0.495, SD = 0.251) and spaced (M = 0.7, SD = 0.228) five-step procedures ($t(34) = -3.445$, $p = 0.001$, $BF_{10} = 29.055$).

For participants in the massed condition, there was a significant difference between the two and five-step procedures ($t(34) = 3.214$, $p = 0.003$, $BF_{10} = 12.466$) after Bonferroni adjustment and for participants in the spaced condition there was also a significant difference between the procedures ($t(31) = 4.107$, $p < 0.001$, $BF_{10} = 100.593$), after Bonferroni adjustment.

Figure 3-4

Experiment Two - Main Effect of Spacing



Note. A boxplot showing the percentage score on the post-test by spacing condition and number of steps in the procedure. The significance stars represent: * p < .05, ** p < .01, *** p < .001

Table 3-7

Experiment two breakdowns of post-test subtasks

Spacing	Procedure	Basic	Correct (y/n)	Recognise Steps	Reverse Procedure
Massed	Two-step	0.525 (0.266)	0.755 (0.263)	0.814 (0.309)	0.676 (0.676)

Spacing	Procedure	Basic	Correct (y/n)	Recognise Steps	Reverse Procedure
	Five-step	0.363 (0.291)	0.647 (0.306)	0.732 (0.311)	0.373 (0.373)
Spaced	Two-step	0.681 (0.190)	0.922 (0.165)	0.902 (0.225)	0.863 (0.863)
	Five-step	0.578 (0.291)	0.804 (0.219)	0.904 (0.246)	0.637 (0.637)

Note. Table displaying the mean and standard deviation for the percentage score on the post-test, broken down by the sub-tasks

3.2.6 Exploratory analyses

Again, we ran four t-tests, with Bonferroni correction, to test if there was a significant difference of working memory, arithmetic fluency, mathematics anxiety, or accuracy across the two retrieval tasks between the massed and spaced conditions. We found no significant difference between any of the variables after Bonferroni correction (see Table 3-8), however, arithmetic fluency would be significant without this correction.

Table 3-8

Experiment two exploratory variables t-tests

	Statistic	d.f.	p	p.adj
Arithmetic Fluency	-2.426	61.692	0.018	0.073
Mathematics Anxiety	1.663	64.336	0.101	0.404

	Statistic	d.f.	p	p.adj
Retrieval Accuracy	-0.778	65.868	0.440	1.000
Working Memory	-0.904	65.930	0.369	1.000

Note. Table displaying the results of t-tests to investigate if there are any significant differences between massed and spaced groups for experiment two. The p values are adjusted using the Bonferroni correction.

3.2.7 Discussion

In the second experiment, we once again tested how manipulating the number of steps in a procedure affected the efficacy of spaced retrieval practice. As in Experiment one, there was a significant effect of spacing, however in contrast, post-hoc tests, done for completeness, showed that this effect was significant across both complexity conditions. We compared a two-step procedure to a five-step procedure. Unlike in Experiment one, we saw a significant main effect of complexity, with worse performance in the higher complexity condition. Increasing the number of steps from three to five in the high complexity condition therefore had a demonstrable effect on performance, providing us with greater certainty that we appropriately manipulated procedural complexity. Critically, despite clear evidence for a main effect of complexity, we again found no evidence for an interaction between spacing and complexity.

3.3 General discussion

We investigated the relationship between procedural complexity and the spacing effect for learning mathematics procedures. In both experiments, participants learnt two arithmetic procedures over either a massed or spaced schedule. We found clear evidence for a main effect

of spacing in both experiments, and post-hoc tests revealed a significant spacing effect in three out of four procedures across the experiments. Critically, we saw no evidence of an interaction between spacing and procedural complexity: the spacing effect was not modulated by complexity. This was also the case in experiment two where a clear main effect of complexity was present. Our results provide further evidence of the robust nature of the spacing effect and support the existing literature recommending the use of spaced retrieval practice in mathematics learning (Emeny et al., 2021; Lyle et al., 2020; Rohrer & Taylor, 2006). Critically, the spacing effect appears to be robust to changes in procedural complexity.

Our first hypothesis was that spaced retrieval would lead to greater retention than massed retrieval. We found a significant main effect of spacing in both experiments, this suggests that the change in scheduling condition from a single massed session to three sessions (spread over three days) provided a significant boost to recall after a week's delay. Importantly, participants performed the same number of practice trials, the only difference was when they performed the practice.

Although we found evidence for a main effect of spacing across both experiments, the post-hoc tests, performed for completeness, suggested the three-step procedure (experiment one) did not show a significant spacing effect (and the Bayes Factor was inconclusive). Although this post-hoc test could simply be a false negative, the deficient processing account (Hintzman, 1974) could explain the lack of spacing effect for the three-step procedure. If the three-step procedure required additional processing it may have not been negatively affected by the deficient processing in massed practice. The two-step procedure may have been simple enough that processing dropped over a small number of trials, therefore those participants benefited from the boost to attention provided by spaced practice. However, if that were the case then we would

have expected a similar effect for the five-step procedure. One possibility is that the five-step procedure may have had too many elements to be processed at once, therefore the lack of deficient processing in the massed condition may not have had an effect. Overall, there is little evidence that this is the case, a false negative is the simpler explanation, particularly given the inconclusive Bayes Factor, the presence of a spacing effect in the other three conditions, and a main effect of spacing in experiment one.

Our second hypothesis was that participants would have lower retention of high complexity material. Experiment one found no evidence for this. In experiment two, performance in the five-step procedure was significantly worse when compared to the two-step, yielding a significant main effect of complexity. One additional step might not be a large enough difference in procedural complexity to elicit an effect of complexity (in experiment one), while five steps (in experiment two) might begin to push the limits of working memory (Miller, 1956), as the number of chunks available in working memory capacity is often estimated to be four or five chunks (Reynolds et al., 2022). For the five-step procedure, participants had to learn the first step plus the operand and operator for four more steps for a total of nine elements.

Our third hypothesis was that there would be an interaction between spacing and complexity. More specifically, that there would be a smaller effect of spacing in the high complexity condition than in the low complexity condition. We had two reasons to predict this. Firstly, considering the study-phase retrieval hypothesis, longer procedures may be more difficult to retrieve. More difficult retrievals could in turn lead to fewer or more inaccurate retrievals of the original study-phase memory trace, negating any benefit of spacing. Secondly, considering the deficient processing account, we hypothesized that adding additional steps to a procedure may increase the level of processing required to retrieve the procedure. Given a finite

level of maximum potential processing, this may mean there is no beneficial effect to memory of additional processing in the spaced condition. We found no evidence for an interaction between procedural complexity and the spacing effect in either experiment. Furthermore, the results of the Bayesian ANOVA suggest that there was strong evidence for the null hypothesis, that there was no interaction between spacing and procedural complexity. This provides evidence against our initial hypothesis and suggests that spacing can be effective for varying levels of procedural complexity: when operationalised as the number of steps in a procedure. While this goes against our initial hypothesis it suggests that the spacing effect is robust to variations in procedural complexity and supports its use in the teaching and learning of mathematics.

If the number of steps in a procedure alone does not affect the efficacy of spaced retrieval practice, an alternative is to consider the cognitive load theory measure *element interactivity* (Sweller, 1988). This concept makes the important distinction between how many elements one must recall while completing a task and how many must be held in working memory simultaneously during the task. Prior knowledge is key as what may be multiple elements to one learner could be chunked into a single element in a more experienced learner (Chen et al., 2023). Chen et al. (2024) investigated the relationship between element interactivity and the spacing effect. They found that materials higher in element interactivity deplete working memory resources, which can be restored after a short rest. This may be one mechanism of the spacing effect seen across short rest periods. In contrast, they also found that for low element interactivity material there is no effect of working memory resource depletion, but they still found a spacing effect, suggesting that working memory resource depletion cannot account for all spacing effects. They suggest that when the spacing effect is not arising from opportunities to restore working memory resources, the effect is due to additional time for rehearsal. In two

further experiments, they taught participants (either novices or more experienced learners) a calculus rule. For novices, for whom they hypothesized the material would be high in element interactivity, they found an effect of spacing and working memory resource depletion. In contrast, for more experienced learners (the material was low in element interactivity) there was no effect of working memory resource depletion or spacing. This suggests that spacing could be effective for material low in element interactivity, through additional rehearsal, and high in element interactivity, through recovery of working memory resources via rest, but perhaps not in material too complicated to rehearse, but that does not deplete working memory resources because of expertise.

The definition used by Sweller (1988) and Chen et al. (2024) suggests that our material was low in element interactivity. While we increased the number of elements required to be learnt to successfully apply the procedure, participants never have to consider all the steps simultaneously. In other words, we have low element interactivity. Chen et al. (2023) suggest that to estimate element interactivity one must calculate the number of elements held in working memory simultaneously. Our participants might not have to hold the whole procedure, but rather, the answer to the previous step, the current step (the operator and operand - two elements) and the answer to the current step in working memory. This would suggest that our maximum element interactivity was four rather than nine. Critically, the number of elements held in working memory at any one time would be consistent across our 2, 3 and 5 step procedures. This could be why our measure of complexity, altering the number of steps, found no significant interaction between procedural complexity and spacing as the element interactivity remains constant, regardless of the number of steps. While our manipulation of procedural complexity decreased performance in the five-step condition, it may not have targeted the aspect of

complexity that affects the efficacy of spacing. Future experiments should develop the material such that participants must hold more information in their working memory simultaneously.

We only used one spacing schedule in both experiments, which reduces generalisability. This experiment used two inter-session intervals of one day, before a retrieval interval of one week. We found a spacing effect for the highest complexity material, so, perhaps our is well optimised for the higher complexity material. If increasing the inter-session interval increased the difficulty of retrieval, then perhaps this would mean that the lower complexity material is still retrieved, while the higher complexity is not. In this case, according to the study-phase retrieval account, the material would not benefit from spacing as the original study-phase would not be retrieved. Therefore, there would be no benefit of spacing. Similarly, as prior studies found that the spacing effect was greatest for medium performers (Barzagar Nazari & Ebersbach, 2019), longer inter-session intervals could reduce the performance of the more complex procedure. This would in turn lower the success rate and may result in a lack of spacing effect.

The main practical takeaway from this study is that if a procedure can be taken step by step, spacing can be effective for procedures of varying complexity. This is useful in cases where it is not meaningful to break a task into smaller sub-tasks. The spacing effect is a powerful and robust way to boost retention. Teachers should ensure students are able to practice material over multiple sessions, rather than massed into a single session in order to maximise learning.

3.4 Conclusion

Across two experiments, we found significant main effects of spacing. Critically, the spacing effect was not modulated by procedural complexity. Our results suggest that, when defined as the number of steps in a procedure, procedural complexity does not affect the efficacy of spaced retrieval practice for mathematics. This provides further support to the robust nature of

the spacing effect that makes it a valuable tool to improve memory and learning. Future research should focus on the structure of procedures, and how the steps interact, rather than the quantity of steps required to follow the procedure to investigate the relationship between procedural complexity and the spacing effect.

4 Element Interactivity and the Spacing Effect

Additional information for journal style thesis:

The following chapter has not yet been submitted at a journal; however, it is presented in a manuscript style to maintain consistency across experimental chapters.

Statement of authorship:

This paper was authored by Ewan Murray and reviewed, edited, and supervised by Aidan J. Horner and Silke M. Göbel. Ewan Murray performed all data analysis, and the results were reviewed by Aidan J. Horner and Silke M. Göbel.

Data availability:

The data and code required to reproduce this manuscript, and all analyses is available on the OSF (***https://osf.io/t3yrh/***).

Abstract

Across two experiments I investigated whether element interactivity, the maximum number of items held in working memory at once during a task, affected the efficacy of spaced practice. The spacing schedule matched the one used in the chapter 3 experiments: either one massed session or three consecutive spaced sessions followed by a seven-day delay. In experiment three, I compared a linear category (low element interactivity) to a relational category (higher element interactivity). There was a significant difference in accuracy on the post-test for the low versus higher element interactivity material, however, no main effect of spacing or interaction. This was the first time I did not find a main effect of spacing during this PhD project. In experiment four I compared the linear category versus the procedure used in chapter 3 and participants completed either massed or spaced practice. This time there was a main effect of spacing, but no main effect of task nor an interaction between spacing and task. Both the procedure and category tasks benefitted from spacing. This suggested that there may have been high levels of interference between the two category types which masked the effect of spacing in Experiment 3. Neither experiment provided evidence that element interactivity moderates the spacing effect.

Keywords: Mathematics learning, Memory, Spacing effect, Distributed practice, Complexity, Element Interactivity

Element Interactivity and the Spacing Effect

The two experiments presented in this chapter were designed to further investigate what aspects of task complexity affect spacing, in particular I focus on element interactivity. An overview of spacing and complexity was provided in Chapter 1 and briefly in chapters 2 and 3, in this next section I focused on element interactivity and how it differs from our previous measure of complexity.

Previously, I manipulated procedural complexity by varying the number of steps in an arithmetic procedure and participants completed either a massed or spaced practice schedule. I proposed that more steps meant higher procedural complexity. I found a main effect of spacing in both the two- versus three-step procedures and two- versus five-step procedure experiments, however there was no evidence of an interaction between spacing and procedural complexity. I concluded that the number of steps did not affect the efficacy of spaced practice. This may be because while learning a procedure participants can learn one step at a time and, once they have recalled and applied the step, they do not need to maintain it in working memory. Therefore, it is possible that when a task can be broken down easily into subtasks, the number of subtasks does not interact with the efficacy of spacing. In contrast to procedural complexity another suggested measure of task complexity, *element interactivity*, may affect spaced practice as it provides an alternative way to think about complexity as opposed to just more items to remember. *Element interactivity* is defined as the maximum number of elements a learner has to hold in working memory simultaneously in order to accomplish a task (Chen et al., 2023). Recently, *element interactivity* has been more clearly operationalised as a measurement of task complexity (Chen et al., 2023), after previous critiques suggested it was not sufficiently well defined (Karpicke & Aue, 2015).

As described in full in chapter 1, Chen et al. (2024) suggest that element interactivity moderates the spacing effect. They propose that higher element interactivity material depletes working memory resources leading to a reduced quality of learning during massed practice and that spacing allows for working memory resources to be recovered during rest periods. However, they also found a spacing effect for tasks that did not deplete working memory resources, in which case they suggest this was because participants were able to rehearse. They manipulated element interactivity by comparing a novice group to an expert group on the same task, suggesting that the expert group's prior knowledge means the material had lower element interactivity. In contrast, I aimed to improve on this design by manipulating element interactivity within subject using artificial material where prior knowledge should have a minimal effect and what I manipulated were the connections between the elements of the task.

In experiment 3, I manipulated the structure of categories. In contrast to the procedures task, when categorising a number, participants need to retain which steps they have previously checked alongside the current step they are checking, this may add an increased load to working memory and increase the interactivity between elements participants were required to learn. In the current study, using element interactivity as the complexity manipulation, I hypothesised that spacing would be less effective as element interactivity increased. I chose this hypothesis as Chen et al. (2024) found conflicting results, where spacing was either found for low element interactivity material (where rehearsal could occur) or higher element interactivity (where spacing provided time for working memory resources to recover). In light of these conflicting results, I considered Donovan and Radosevich (1999) meta-analysis results to be stronger evidence, where overall task complexity was negatively correlated with the magnitude of the spacing effect. I had three hypotheses:

H1) Spaced practice will lead to a greater retention than massed retrieval

H2) Participants will have lower retention of higher element interactivity material

H3) There will be a significantly smaller effect of spacing in the higher element interactivity condition.

4.1 Experiment Three

4.1.1 Method

4.1.2 Participants

As I expected a similar pattern of results to the previous experiments in chapter 3, I reused the simulated power analysis, which suggested I needed 34 participants per group to power for a large effect of spacing typical in the domain of spacing (η^2 = 0.11). This was 2(only for the main effects, however, as recent best practices have suggested much larger sample sizes for interaction effects (Sommet et al., 2023). Furthermore, I hypothesised increased complexity would attenuate the spacing effect, reducing its efficacy, but not making massed practice more effective than spaced practice. Attenuated interactions are harder to power for than a full cross over interaction (Lakens & Caldwell, 2021). The power analysis suggested that 218 participants (109 per group) were required to fully power for the expected interaction. As this material was new and time was limited at this stage in the project, I did not have the resources to fully power for the interaction. The final sample consisted of 80 undergraduate students at the University of York UK. The average age of the participants was 19.78 years (SD = 1.71 years), 68 participants identified as female, eleven as male and one as non-binary.

4.1.3 Procedure

The University of York's human participant pool was used for recruitment via Sona Systems (https://www.sona-systems.com/). Participants accessed the experiment through the

experimental platform Gorilla (Anwyl-Irvine et al., 2020). Participants were granted course credit upon completion of the study. After signing up participants were asked their age and gender then chose when they would complete the study. Both Experiments 3 and 4 were approved by the University of York's Ethics board, and all participants gave explicit consent to participate.

As I found large spacing effects across both experiments 1 and 2 (see chapter 3) I used the same spaced practice schedule and only changed the material participants were tasked to learn. All participants who completed the experiment participated over four days: three consecutive and one session following a seven-day gap. The spacing manipulation was applied to the category learning task, and the individual differences tasks were used to pad the remaining time. As this task took longer than the main task in the Experiments 1 and 2, I omitted the maths anxiety test in Experiment 3.

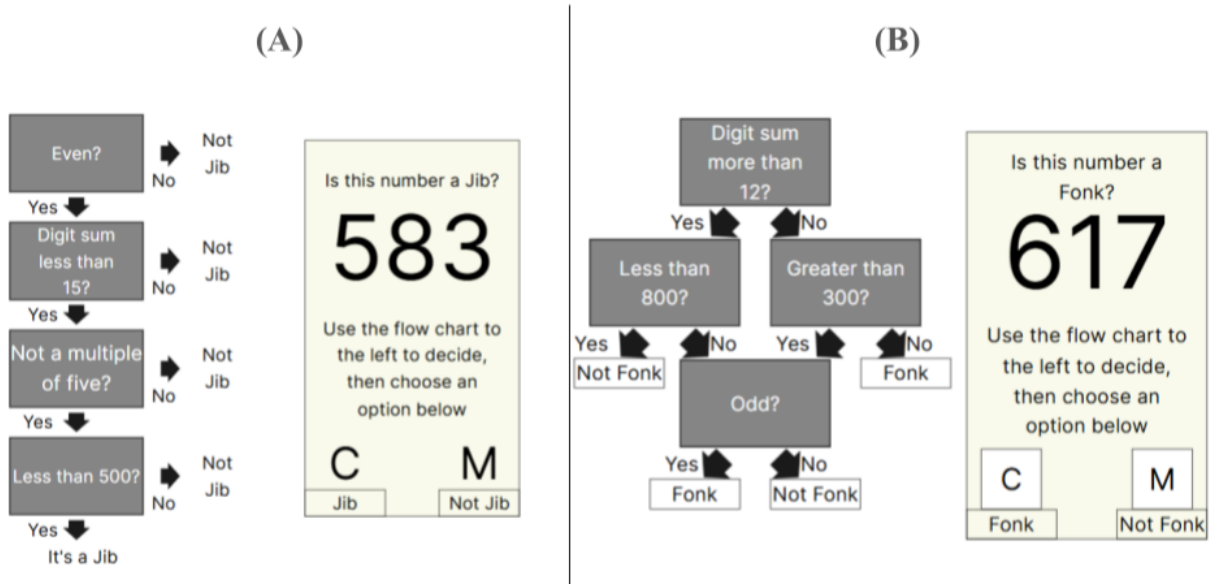
4.1.4 Material

Explicit Category Learning. During the main task, each participant learnt to categorise a number using one linear category (lower element interactivity) and one relational category (higher element interactivity) (see Figure 4-1). Linear categories required participants to recall four constraints and, as long as all the constraints were true, the number was a member of the category (see Figure 4-1 **-** A). As participants were able to learn each constraint in isolation, this task was designed to have low element interactivity. Though participants were always shown the flowchart to categorise a number in the same order, the actual order they checked the constraints has no bearing mathematically. Additionally, when the target number was not a member of the category, to successfully apply the *linear* condition participants only needed to spot one constraint that did not hold (i.e., that the number was less than 500) at which point they could

respond that it was not a category member. To be sure that it was a category member they needed to check all four constraints.

Figure 4-1

Screens Shown During the Learning Phase



Note. Screens participants saw during the learning of linear (A) and relational (B) tasks.

In contrast, to apply the relational category simply checking all the steps was not sufficient. Each constraint was only useful when performed in the correct order and with knowledge of what the participant has previously checked. Due to the structure of the category (see Figure 4-1 - B), successful recall of any one constraint for the relational category would not allow participants to accurately categorise a number. This structure increased the number of elements required to be held in working memory at once, in turn making the relational category have higher element interactivity.

I ran a pilot study to check that each constraint had a similar response time, using this as a measure of difficulty and participants' prior knowledge. I found only very small differences in

response times between the constraints, except for working out the sum of the digits in a number, which took longer for participants to respond to. To minimise this issue, each category included either “Digit sum more than 10” or “Digit sum less than 15”. See Table 4-1 for the full list of constraints.

Table 4-1

Constraints used to form categories

Set one	Set two
Digit sum more than 12?	Digit sum less than 15?
Less than 800?	Even?
Greater than 300?	Not a multiple of five?
Odd?	Less than 500?

Note. Table displaying the evaluative statements used to form the numerical categories, participants learned both sets, one in a linear structure one in a relational structure.

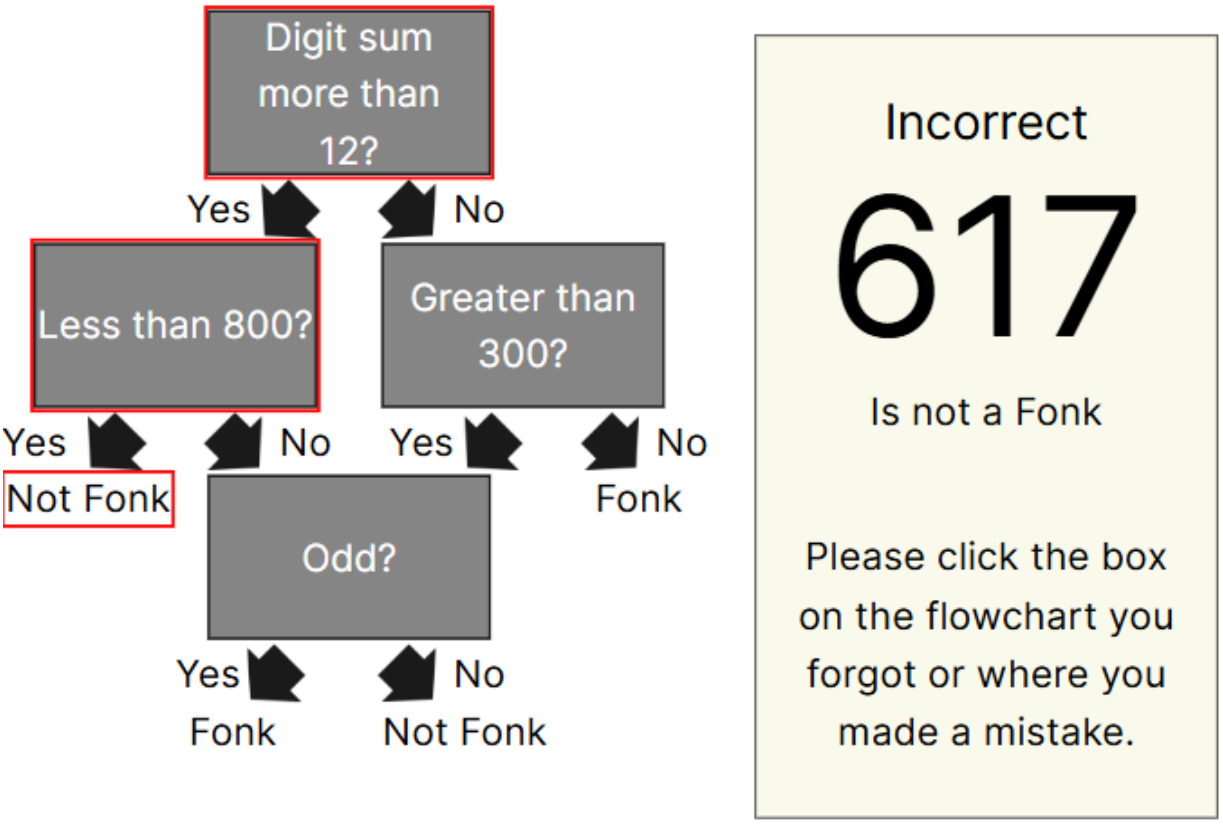
For the main task, I developed the categories to make them as similar as possible. First, it was necessary to ensure that half of the trials were members of the category and half were not, because participants either pressed a key to say the trial was a member of the category or another key to say that it was not. Therefore, answering randomly or only pressing one key throughout practice would result in 50% accuracy and I could maintain a 50% chance for both category types. Another important consideration was ensuring that if participants were performing the task as instructed, they would reach each potential end condition of the flowchart evenly. In the main

task, across all trials participants checked an average of 3.25 constraints for the linear category and 2.5 constraints for the relational category, if they were performing the task as instructed.

During the practice trials participants saw the flowchart alongside the number they were required to check against (see Figure 4-1). They then had to press C or M to categorise the number as either a member of the category or not. During each session participants completed 26 trials. During the learning and retrieval phases of the experiment when participants got an answer incorrect they were asked to click where on the flowchart they made an error or forgot the constraint (see Figure 4-2). I chose to do this as immediate feedback is a potential boundary condition of the spacing effect (Emeny et al., 2021). The main task of the post-test consisted of 24 additional trials, identical to practice, but with no feedback. Two additional tests were added for exploratory purposes, however, they provided no insights, so are only reported in the appendix (see appendix A).

Figure 4-2

Feedback Shown When the Answer was Incorrect



Note. Example of feedback seen by participants when incorrectly categorising a number in the relational task.

To enable comparisons with our previous experimental work (see chapter 3) I used the same *working memory*, *arithmetic fluency*, and *mathematics anxiety* tasks to measure participants' individual differences. I briefly summarised them again, but please see chapter 3 for a full description. To measure participants' working memory I employed a forwards and backwards digit span task based on the Wechsler memory scale - third edition (WMS-III) (Wechsler, 1997). Arithmetic fluency was measured using the Math4Speed task (Loenneker et

al., 2022) and participants answered as many addition, subtraction, multiplication, and division questions as possible (two minutes per arithmetic type). Mathematics anxiety was measured using the Mathematics Anxiety Scale–UK (MAS-UK) (Hunt et al., 2011), a scale designed to measure mathematics anxiety for undergraduate students in the UK.

4.1.5 Analysis

I preregistered the analysis plan (***https://osf.io/yf4nm***) and planned to run a two (spacing: massed versus spaced) by two (complexity: lower versus higher element interactivity) between-within subject Mixed ANOVA on accuracy on the post-test. I planned to run the analysis using both frequentist hypotheses testing methods and using the BayesFactor R package (Morey & Rouder, 2023) with the default Jeffreys priors to calculate a Bayes Factor for each effect. The inference criteria for the frequentist statistics were that the p-value must be less than 0.05 for the effect to be significant.

4.1.6 Results

4.1.7 Descriptive Statistics

This experiment had an attrition rate of 32%. Initially, 117 participants completed the first session of the experiment and 80 finished overall. Twenty participants in the spaced condition did not finish, while seventeen participants in the massed condition did not finish.

Table 4-2

Experiment three descriptive statistics

Spacing	Complexity	Overall Score	Retrieval Accuracy	Working Memory	Arithmetic Fluency
Massed	Linear	0.765 (0.226)	0.872 (0.143)	20.732 (3.860)	78.366 (25.483)
	Relational	0.639 (0.203)	0.737 (0.172)		
Spaced	Linear	0.762 (0.231)	0.857 (0.136)	19.564 (4.309)	74.308 (24.628)
	Relational	0.629 (0.216)	0.744 (0.148)		

Note. Table displaying the mean and standard deviation for the percentage score on the post-test, retrieval accuracy (the mean percentage accuracy across both practice retrieval sessions) and the scores on the working memory, and arithmetic fluency tasks

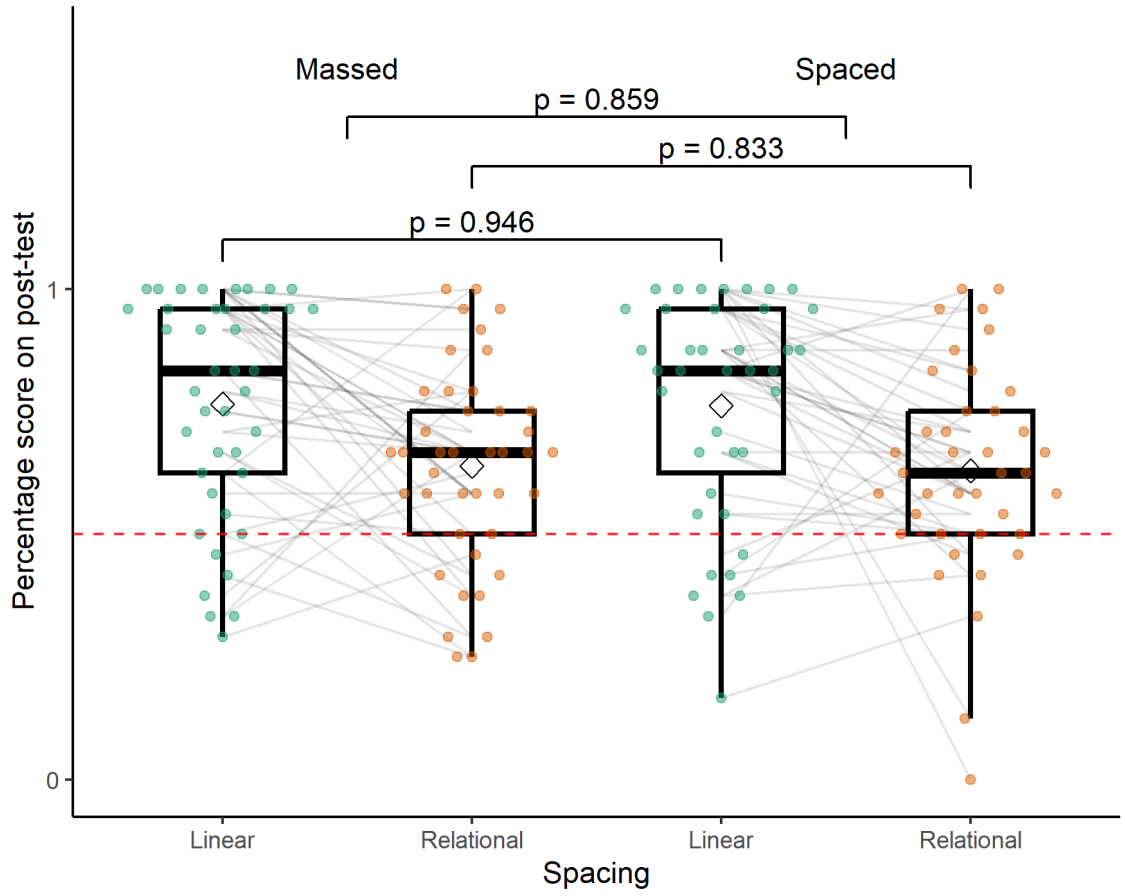
4.1.8 Main Analysis

A two (spaced/massed) by two (linear/relational) mixed ANOVA was performed, with percentage accuracy on the post-test as the dependent variable (see Figure 4-3). There was no significant main effect of spacing, $F(1,78) = 0.032$, $p = 0.859$, $\eta^2 < 0.001$ ($BF_{10} = 0.222$). The percentage overall score on the post-test was not significantly different in the spaced ($M = 0.696$,

SD = 0.232) than the massed condition (*M* = 0.702, *SD* = 0.223). There was a significant main effect of complexity, $F(1,78) = 17.180$, $p < 0.001$, $\eta^2 = 0.082$ ($BF_{10} = 483.040$). The percentage overall score on the post-test was significantly higher in the Linear (*M* = 0.764, *SD* = 0.227) than the Relational condition (*M* = 0.634, *SD* = 0.208). There was no evidence of an interaction between spacing and complexity, $F(1,78) = 0.011$, $p = 0.918$, $\eta^2 < 0.001$ ($BF_{10} = 0.413$).

Figure 4-3

Experiment Three: Post-test Performance After Massed versus Spaced Practice by Complexity Condition



Note. A boxplot showing the percentage score on the post-test by spacing condition and complexity (linear versus relational category). The dashed line denotes chance at 50%.

4.1.9 Exploratory Analysis

Individual difference measures

There were no significant differences between the two groups that may have biased the final results (see Table 4-3).

Table 4-3

Experiment three exploratory variables t-test

	Statistic	d.f.	p	p.adj
Arithmetic Fluency	0.72	77.98	0.47	1.00
Retrieval Accuracy	0.47	78.00	0.64	1.00
Working Memory	1.27	76.06	0.21	0.62

Note. Table displaying the results of t-tests to investigate if there were any significant differences between massed and spaced groups for experiment three. The p values were adjusted using the Bonferroni correction.

4.1.10 Discussion

I found no evidence of a spacing effect. This contrasts with the results in chapter 3 where participants learnt procedures and benefited from an identical spaced practice schedule. For the linear condition, participants appeared to be performing the task properly. Performance was high and above chance. For the relational condition, participants may have not been performing the task properly or may not have understood the task's instructions completely. There was a ceiling effect in the linear condition, however the distributions across the spacing condition were very

similar. I would have hypothesised a much stronger ceiling effect in the spaced condition given the massed data.

I also ran exploratory analyses to check for relationships between the accuracy in either the cued recall or recognition tasks(see Appendix XX). No significant effects were seen in these analyses. There were no significant between group differences for working memory, retrieval accuracy or working memory. Exploratory correlation analyses suggested no significant relationship between any individual difference measure and post-test performance, as these analyses provided little theoretical insight they were not included.

If higher element interactivity is a boundary condition of the spacing effect, then one reason I may have not gotten a spacing effect in either condition could be that both the linear and relational categories were too high in element interactivity. Both tasks require the participant to hold constraints in working memory while checking other constraints and could have put too high a load on working memory, in contrast to the procedures used in chapter 3.

Alternatively, the constraints used in the two categories were very similar, which could have led to high levels of interference across the two conditions. In our meta-analysis, isolated material had a larger mean effect of spacing than material embedded in a course, in which recall of one item may have interfered with the recall of another (see chapter 2). If this were the case this could be an important boundary condition of the spacing effect and impact interleaving techniques, which has not previously been considered a boundary condition in the mathematics learning and spacing domain (Emeny et al., 2021).

In experiment four I compared a category learning task to a procedure learning task. If the linear category in experiment 3 was too high in element interactivity for spacing to be effective, then I would expect the linear category to not benefit from spacing when it is learnt

alongside a procedure in experiment 4. In contrast, if I did not get a spacing effect because of high interference then the linear category may benefit from the spacing effect in experiment four.

4.2 Experiment Four

4.2.1 Method

I used an identical experimental design to experiment three, but I changed the material participants were taught and how it was displayed (see Figure 4-4). The reasons for this change are explained in the materials section below. Participants learnt one five-step procedure (adapted from experiment two) and an adapted five-constraint linear category employed in the previous experiment.

4.2.2 Participants

I used the same experimental design as before, so planned to recruit a minimum of 68 participants (34 per condition) to achieve adequate power to detect the main effects but would require more than 200 participants to detect a medium interaction which was not feasible at this stage in the project. The average age of the participants was 27 years old (SD = 9.48 years). Forty-six participants identified as female and twenty-four as male. Participants were recruited through prolific and gave informed consent.

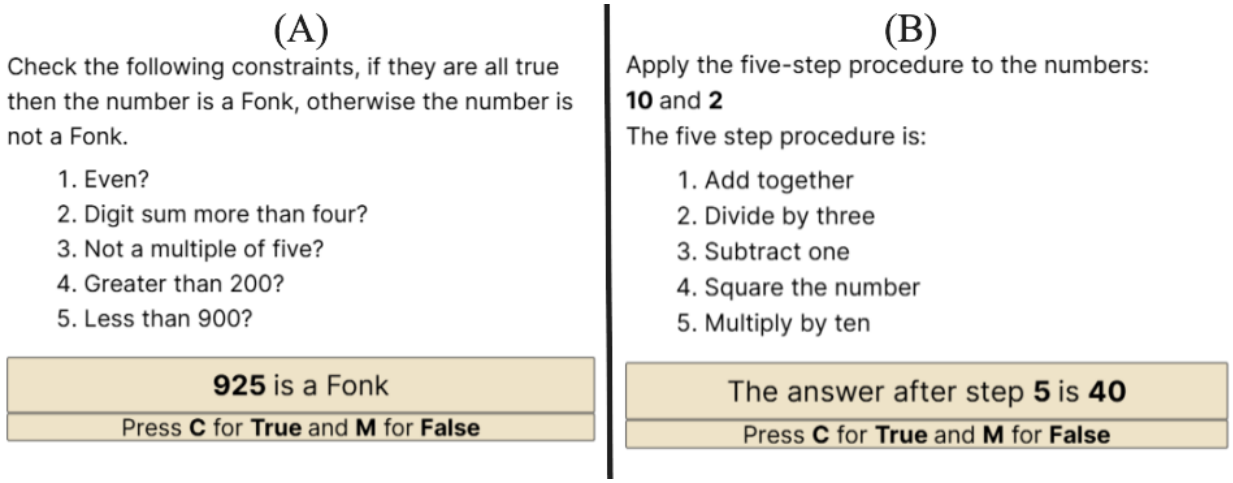
4.2.3 Material

Explicit Category Learning

I moved from a four-constraint category to a five-constraint category to make the task more difficult, to reduce performance, and enable any spacing effect to be viewed more clearly. In experiment three, for the four-constraint category each session had 24 trials, however, in experiment four the five-constraint category had 20 trials. I expected that decreasing the number of practice trials would reduce performance in the category condition.

Figure 4-4

Experiment Four: Learning Phase



Note. Examples of the screens participants may have seen during the learning phase. In the retrieval phases and post-test participants would see the same screen without the steps/constraints available

Learning arithmetic procedures

I used the five-step procedure from experiment two (see Figure 4-4 - A). It required participants to learn to apply five arithmetic steps to a pair of numbers. During the experiments in chapter 3, participants answered each trial by typing their answer as a number. As the category learning task has a chance level of 50%, I wanted to change how participants answered the learning arithmetic procedures task to reflect that. In the current experiment I provided participants with a number, which was either the correct or incorrect answer, and participants had to provide a binary response whether it was the correct answer or not. This change allowed for both tasks to consist of a binary choice and minimised the differences between the two tasks (see Figure 4-4).

Participants were asked whether the answer was correct after a certain step or after the completion of all the steps. The number of steps and constraints needed to be checked was matched across the two tasks. This allowed us to match the number of steps required to check the answer to the number of constraints participants would be required to check (if they followed the checking procedure in the order provided) and use that to classify the number without having to check in the original order.

4.2.4 Results

This experiment had an attrition rate of 23%. Initially, 90 participants completed the first session of the experiment and 69 finished overall. Twelve participants in the spaced condition did not finish, while nine participants in the massed condition did not finish.

4.2.5 Descriptive Statistics

Table 4-4

Experiment Four Descriptive Statistics

Spacing	Task	Overall Score	Retrieval Accuracy	Working Memory	Mathematics Anxiety	Arithmetic Fluency
Massed	Procedure	0.858 (0.102)	0.921 (0.069)	19.800 (5.556)	46.743 (18.454)	82.886 (27.931)
	Category	0.792 (0.2)	0.893 (0.142)			
Spaced	Procedure	0.913 (0.083)	0.832 (0.169)	22.057 (4.518)	49.829 (17.154)	78.857 (31.746)

Spacing	Task	Overall Score	Retrieval Accuracy	Working Memory	Mathematics Anxiety	Arithmetic Fluency
	Category	0.913 (0.121)	0.861 (0.098)			

Note. Table displaying the mean and standard deviation for the percentage score on the post-test, retrieval accuracy (the mean percentage accuracy across both practice retrieval sessions) and the scores on the working memory, mathematics anxiety, and arithmetic fluency tasks

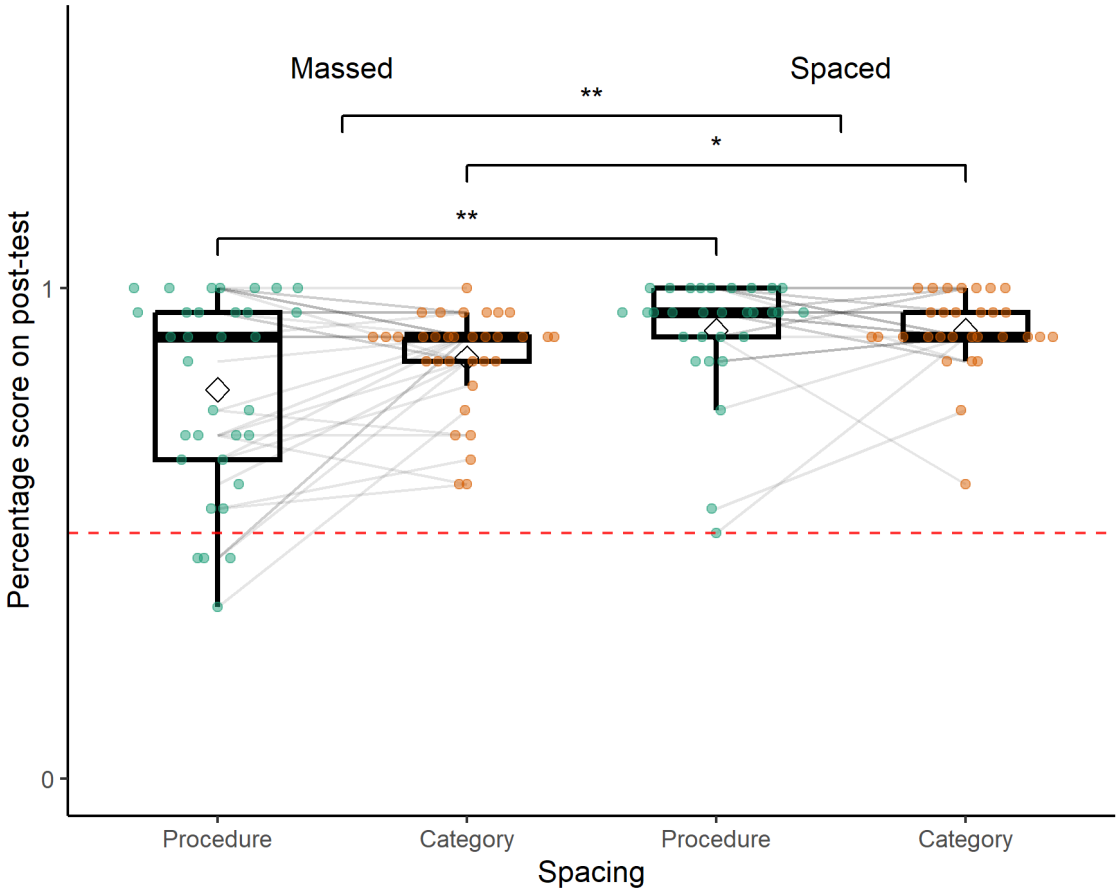
4.2.6 Main Analysis

When I analysed post-test performance seven participants (five spaced, two massed) were identified as extreme outliers with low performance (SD > 3) and were removed as planned in the pre-registration. Additionally, there were concerns with heterogeneity of variance and normality of the underlying distribution. To address these concerns, I performed the planned standard mixed ANOVA (with outliers removed), followed by a robust ANOVA on the trimmed means (using the total dataset), which as an analysis is more robust to a lack of normality, heterogeneity, and outliers. First, I performed a two (spaced/massed) by two (category/procedure) mixed ANOVA, with percentage accuracy on the post-test as the dependent variable (see Figure 4-5). There was a significant main effect of spacing, $F(1,61) = 9.578$, $p = 0.003$, $\eta^2 = 0.099$ ($BF_{10} = 11.617$). The percentage overall score on the post-test was significantly higher in the spaced condition ($M = 0.913$, $SD = 0.103$) than the massed condition ($M = 0.825$, $SD = 0.161$). There was no significant main effect of task, $F(1,61) = 2.997$, $p = 0.088$, $\eta^2 = 0.015$ ($BF_{10} = 0.815$). The percentage overall score on the post-test was numerically lower in the category ($M = 0.850$, $SD = 0.177$) than the procedure condition ($M = 0.884$, $SD =$

0.097). There was no significant interaction between spacing and task, $F(1,61) = 2.997$, $p = 0.088$, $\eta^2 = 0.015$ ($BF_{10} = 0.920$). The model with the strongest Bayesian evidence was the model that included both the Spacing and Spacing:Task interaction ($BF_{10} = 12.067$).

Figure 4-5

Experiment Four: Post-test Performance After Massed versus Spaced Practice by Task Condition



Note. A boxplot showing the percentage score on the post-test by spacing and task conditions. The dashed line denotes chance at 50%. The significance stars represent: * p < .05, ** p < .01, *** p < .001

Second, I performed the Robust ANOVA using the WRS2 (Wilcox, 2012) R package. These results mirrored the traditional ANOVA, showing a significant main effect of Spacing (see Table 4-5).

Table 4-5

Experiment Four - Robust ANOVA

	df1	df2	Q	p
Spacing	1	32.066	4.650	0.039
Task	1	31.790	1.458	0.236
Interaction	1	31.790	3.376	0.076

Note. Table displaying the results of the robust ANOVA using 20% trimmed means via the WRS2 R package

4.2.7 Individual Difference Measures

I again tested for differences between the two groups for the individual difference measures. In contrast to experiment three, retrieval accuracy was significantly higher in the massed group during practice than the spaced group (see Table 4-6).

Table 4-6

Experiment Four Exploratory Variables t-test

	Statistic	d.f.	p	p.adj
Arithmetic Fluency	0.56	66.92	0.57	1.00

	Statistic	d.f.	p	p.adj
Mathematics Anxiety	-0.72	67.64	0.47	1.00
Retrieval Accuracy	2.90	44.95	0.01	0.02
Working Memory	-1.86	65.28	0.07	0.27

Note. Table displaying the results of t-tests to investigate if there were any significant differences between massed and spaced groups for experiment four. The p values were adjusted using the Bonferroni correction.

4.2.8 Follow-up Analysis

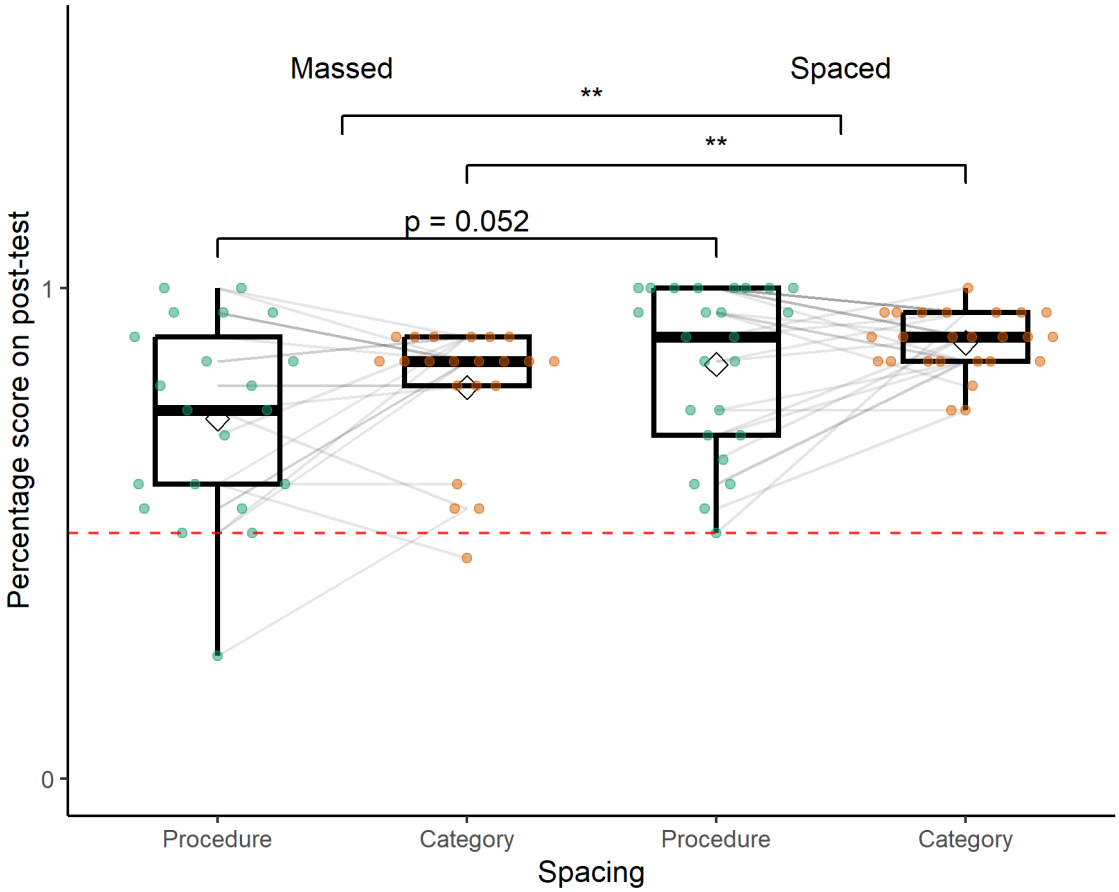
Due to the ceiling effect in experiment four I decided to gain ethical approval to run a follow up study to track performance after a longer timescale. Two weeks after each participant finished the initial post-test they were presented with another post-test with the exact same procedure, but different numbers. Participants were told that it was optional and that they did not have to take part. Twenty-one participants originally assigned to the massed condition returned alongside twenty-five participants in the spaced condition. This group of participants had a mean age of 28.48 years (SD = 10.38) and twelve of the participants identified as male.

I ran the same analyses as I did for the initial post-test. A two (spaced/massed) by two (category/procedure) mixed ANOVA was performed, with percentage accuracy on the post-test as the dependent variable (see Figure 4-6). There was a significant main effect of spacing, $F(1,44) = 7.908$, $p = 0.007$, $\eta^2 = 0.106$ ($BF_{10} = 8.057$). The percentage overall score on the post-test was significantly higher in the spaced ($M = 0.866$, $SD = 0.127$) than the massed condition ($M = 0.765$, $SD = 0.173$). There was a significant main effect of task, $F(1,44) = 4.470$, $p = 0.040$, η^2

= 0.033 (BF_{10} = 1.654). The percentage overall score on the post-test was significantly lower in the category (M = 0.793, SD = 0.190) than the procedure condition (M = 0.847, SD = 0.111). There was no evidence of an interaction between spacing and task, $F(1,44) = 0.157$, $p = 0.694$, $\eta^2 = 0.001$ (BF_{10} = 1.541).

Figure 4-6

Procedure Versus Category - Main Effect of Spacing (Two Weeks Later)



Note. A boxplot showing the percentage score on the post-test by spacing condition and number of steps in the procedure. The dashed line denotes chance at 50%. The significance stars represent: * p < .05, ** p < .01, *** p < .001

Again, the homogeneity of variance assumption to run the ANOVA were not met, so I followed up with a robust ANOVA using the WRS2 R package (see Table 4-7).

Table 4-7

Experiment Four – Follow-up - Robust ANOVA

	df1	df2	Q	p
Spacing	1	23.105	5.189	0.032
Task	1	21.636	2.614	0.120
Interaction	1	21.636	1.529	0.230

Note. Table displaying the results of the robust ANOVA using 20% trimmed means via the WRS2 R package

4.2.9 Discussion

In experiment four, I compared spaced versus massed practice across two tasks. Participants learnt a procedure, similar to the one used in experiment two (with the answer format changed from short answer to whether a provided answer was correct) and a version of the linear category used in the previous experiment (with an extra constraint). I found a significant main effect of spacing, but no significant effect of task nor a significant interaction, however, the interaction between spacing and complexity was closer to significance than in all other experiments in this thesis. Due to monetary, time and resource constraints inherent to a PhD project and this experiment's placement at the end of the PhD project I did not power for the interaction term. This result supports the previous findings that spacing can be a useful technique to improve the learning of mathematics procedures. I first ran the preregistered

analysis; however, two assumptions were violated, homogeneity of variance and the presence of outliers. To minimise the effect of these issues, I ran a robust ANOVA on the trimmed means. These results supported each other, and I have reported them both for transparency. This analysis also aimed to negate the effect of the ceiling effect by removing values at ceiling.

I set prolific up so that participants were supposed to be similarly matched to with the sample in experiment three, however, while all participants claimed to be current UK based undergraduate students, over 18 years of age, the mean age was around 27 years old, in contrast to experiment three, where the participants had a mean age closer to 20. There was also a much larger spread of ages with the standard deviation of 9.48 years for experiment four as opposed to 1.71 for experiment three. It was possible that participants on prolific were more likely to be mature students, however, this may also be an example of inaccurate screening on prolific where participants did not remove their student status after graduation. There is mixed evidence that age may affect the efficacy of the spacing effect, with some studies finding that it doesn't interact (Bercovitz et al., 2017; Kornell et al., 2010) and others finding evidence that age may reduce the magnitude of the spacing effect in verbal materials (Balota et al., 1989; Simone et al., 2013). As I found a spacing effect in both conditions and had high performance this was unlikely to be an issue, however, perhaps I may have found larger spacing effects in a younger population which in turn may have made finding an interaction more likely.

I found that participants' average performance across all practice trials was significantly better in the massed condition than in the spaced condition. This trade-off for lower performance during initial practice that leads to poorer performance after a delay is typical in spaced practice studies (Walsh et al., 2018) , indeed sometimes massed practice is sometimes still more effective for short intervals such as a week instead of four weeks (Rohrer & Taylor, 2006). This

phenomenon was not observed in experiment three or either of chapter 3's studies where practice performance did not significantly differ across groups. Prior studies looking at mathematics and spacing found that the increased fluency participants in the massed group felt during practice led to overconfidence in how well they would perform on a later test, with respect to participants in the spaced condition (Emeny et al., 2021). Participants in the massed group in this experiment could have been overconfident, leading to them paying less attention to the later questions or perhaps rehearsing less once the practice has finished. I found no significant differences between the massed and spaced group for the other individual difference measures: Arithmetic Fluency, Mathematics Anxiety or Working Memory.

The results were hindered by a ceiling effect. I tried to avoid this during development of this material by reducing the number of practice trials and adding an additional constraint to the category learning task, however, it was not sufficient. Alternatively, for the procedure, the ceiling effect could be due to the changes in how the material was presented. In chapter 3 I required a short answer from participants after they completed each trial. This had a much lower chance of making a correct guess. I chose to do this to make the task more comparable to learning the category.

The follow up study, two weeks after the initial post-test, found significant effects of spacing and task, but no interaction. However, this analysis was also potentially unreliable due to heterogeneous variance between conditions. In contrast, the robust ANOVA found a significant main effect of spacing, but no significant effect of task nor a significant interaction. This follow up study was intended to allow for forgetting and reduce performance. Many participants were still close to ceiling, particularly in the spaced condition. Interestingly the main effect of spacing was retained, but analyses suggested that this was no longer significant for the procedure task,

but just for the category task. An important clarification was that the massed condition was no longer truly massed in this case as they took the second post-test two weeks after their first, so both had some spaced practice, albeit without feedback.

4.3 General Discussion

The aim of the two experiments presented in this chapter was to investigate the effect of task complexity on spaced practice in mathematics learning. While the previous experiments in chapter 3 manipulated the number of steps in a procedure, varying the total number of elements required to be recalled, in experiment three and four I sought to manipulate element interactivity. I did this by ensuring that the number of elements needed to be recalled was constant; and instead, varied how the material was structured. Experiment three compared the structure of two different categories, while experiment four compared a category learning task to a procedure learning task.

In experiment three, I taught participants two categories. The rules were available to participants during practice and feedback was provided. I found no effect of spacing in either the linear or relational category condition. This was particularly surprising in the linear condition as, using the same spacing schedule (see chapter 3), I found large effects of spacing for procedures of similar length. This led us to think that perhaps even the lower complexity condition may have been too high in element interactivity for spacing to be effective or that there was too much interference between the two category structures. To investigate whether interference or task complexity was the cause of the lack of a spacing effect I next compared the linear category task to the five-step procedure used in chapter 3.

In experiment four, I compared a five-step procedure (previously used in experiment two) to a five-constraint linear category. I found a significant main spacing effect, but no significant

effect of task nor a significant interaction. The spacing effect was significant in the procedure condition providing further evidence that spacing was effective for learning mathematics procedures, however, there was also a significant spacing effect in the category condition in contrast to experiment three. The overall pattern of the data hints that spacing may have been more effective for the procedure than the category, however, this was not significant and there was no Bayesian evidence to support this. The interaction effect was closer to significance in this experiment than any other during the PhD, suggesting that this may be the best avenue to pursue in the future.

There were three changes to the linear category task from experiment three to experiment four. First, in experiment three the linear category was learnt alongside a relational category task, both tasks used very similar constraints making interference between the two very likely, while in experiment four the steps in the procedure were distinct from the constraints needed to be checked in the category, making interference less likely. Second, the constraints were presented as a list instead of a flowchart. Perhaps the list was easier to integrate as the items were spatially closer together and not in separate boxes. Third, the number of constraints increased from four to five.

There were two factors that may have diminished the ability to find the spacing by task interaction. First, the experiment had low power to detect the interaction. Recent experimental design best practices suggest that to adequately power for an interaction requires hundreds of participants, and I did not predict that the effect would fully reverse but instead merely attenuate which further increases power requirements (Button et al., 2013; Sommet et al., 2023). I did not have the resources to run a fully powered study at this stage of the PhD. Second, the ceiling effects present across both experiments. Despite the attempts to reduce the chance of too high

performance participants were still close to ceiling in all conditions, this may have hidden the interaction effect. As I needed 20 trials per session for category learning counterbalancing, I had 20 trials for the learning arithmetic procedures task, for parity. This was double the number of trials participants completed in chapter 3, however, as participants were no longer required to input a number and instead choose correct or not, I thought the additional trials would provide a better granularity of performance as chance changed from close to zero to 50%. This may however have led to increased performance. In the future, with more time and resources, it would be better to do more extensive piloting to find the minimum number of trials to retain reasonable performance in the post-test, without being at ceiling. Alternatively, practice to criterion could be used. I chose to control the exact amount of practice used in these studies, rather than to criterion, as the deficient processing account of the spacing effect suggested that the amount of practice trials would affect the efficacy of spacing as massed practice would not benefit from further practice once a certain level of performance was reached. I decided to err on the side of too much practice as I believed this would be the best chance to see a spacing effect, considering the deficient processing account (Delaney et al., 2010; Greene, 1989), and capture any interaction. While the ceiling effect caused problems with the statistical analysis there is also some evidence that spacing is less effective for high-performing participants, instead benefiting medium performers the most (Barzagar Nazari & Ebersbach, 2019), there was not however enough evidence to support this in the meta-analysis in chapter 2.

Alternatively, spacing may not interact with task complexity when operationalised as element interactivity. To have more confidence in this claim, further research could focus on reducing interference between the tasks, perhaps by maintaining the structure of the relational versus linear categories but varying the surface features of the task. For example, one category

could categorise a number, as participants were required to do in this study, while the other category could define whether a shape was a member of the group. An example constraint could be whether the shape has a right angle, in contrast to the constraints used in this study such as "is the number even?" This would reduce the effect of interference while retaining the structure of the categories and could be counterbalanced to control for the effect of these surface features. I initially considered implementing this manipulation; however, I considered it to be too far removed from the experiments in chapter 3.

4.4 Conclusion

I tested two novel manipulations of task complexity. Experiment three was the first time I failed to find a main effect of spacing during the PhD project. Neither the linear nor the relational category learning tasks benefited from spaced practice versus massed practice. However, in experiment four the linear category did benefit from spaced practice. The linear category in the second experiment varied slightly from the first as it was designed to better match the procedure task, however the largest difference was that it was learnt alongside a procedure instead of another similar category. In both experiments the complexity or task manipulation did not interact with the effect of spacing, however, experiment four was closer to significance than any other manipulation. Future research should explore the differences between the effectiveness of spacing for low and higher element interactivity material but should ensure that the surface features of the tasks are distinct, limiting the potential interference.

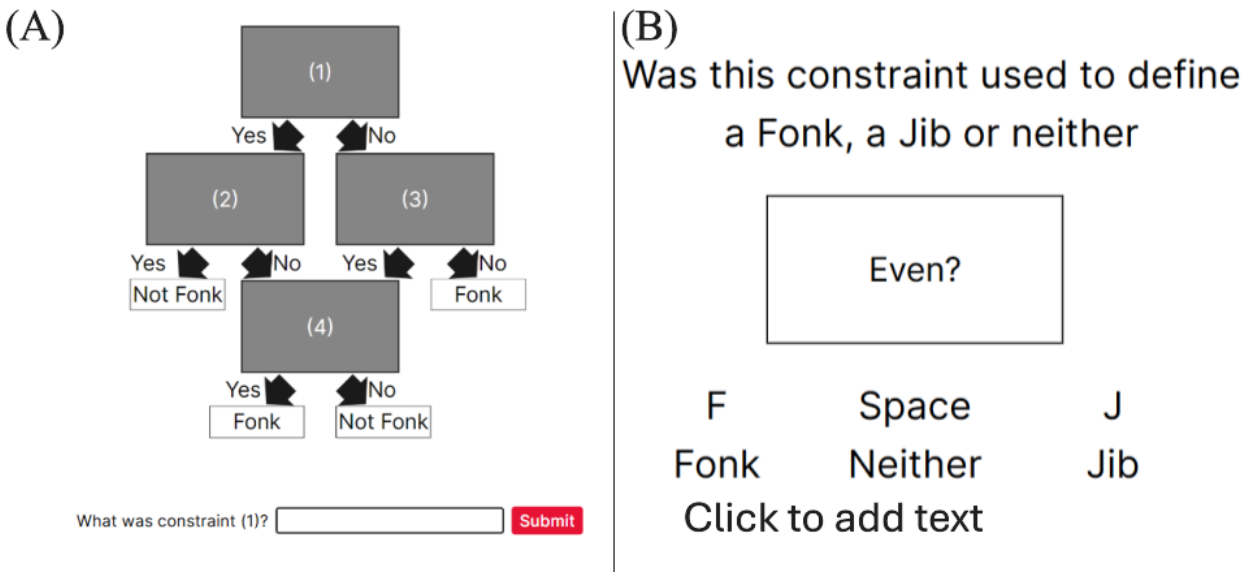
4.5 Appendix

Experiment three Secondary Post-tests

Participants additionally completed a task where they were asked to retrieve each constraint using a labelled flowchart as a cue, the *cued-recall* task (see Figure 4-7- A) and a recognition task which asked them to choose whether a cue belonged to the fonk category, jib category or was a new constraint not used in the task (see Figure 4-7 - B).

Figure 4-7

Participants completed additional trials identical to their practice trials, and a cued recall of constraints task (A) and recognition of each constraint task (B)

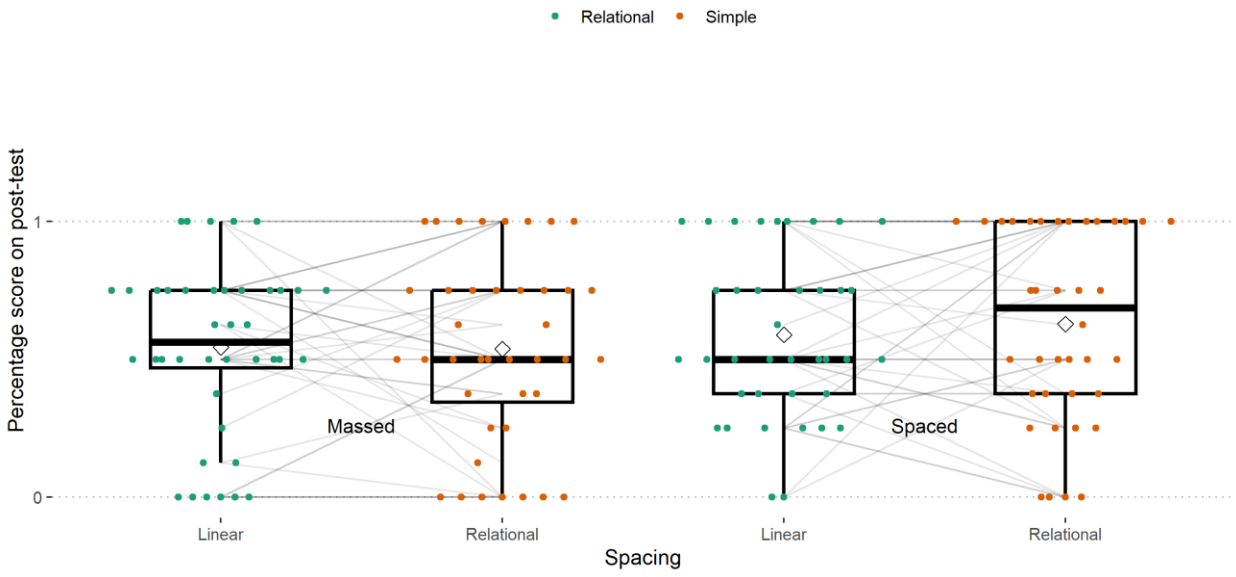


For the cued recall task, participants had to provide a short answer when prompted with either the linear or relational flowchart. Participants were given one point if the constraint belonged to the category and one point if it was in the correct location. Each category consisted of four constraints for a total of 8 points. There was no significant effect of Spacing ($F(1, 76)$ =

1.188, $p = .279$, $\eta^2 = 0.011$, $BF_{10} = 0.406$), Complexity ($F(1, 76) = 0.173$, $p = .678$, $\eta^2 < 0.001$, $BF_{10} = 0.192$) or an interaction between spacing and complexity ($F(1, 76) = 0.328$, $p = .569$, $\eta^2 < 0.001$, $BF_{10} = 0.274$) (see Figure 4-8).

Figure 4-8

Experiment Three: Cued Recall Task Performance After Massed Versus Spaced Practice By Complexity Condition



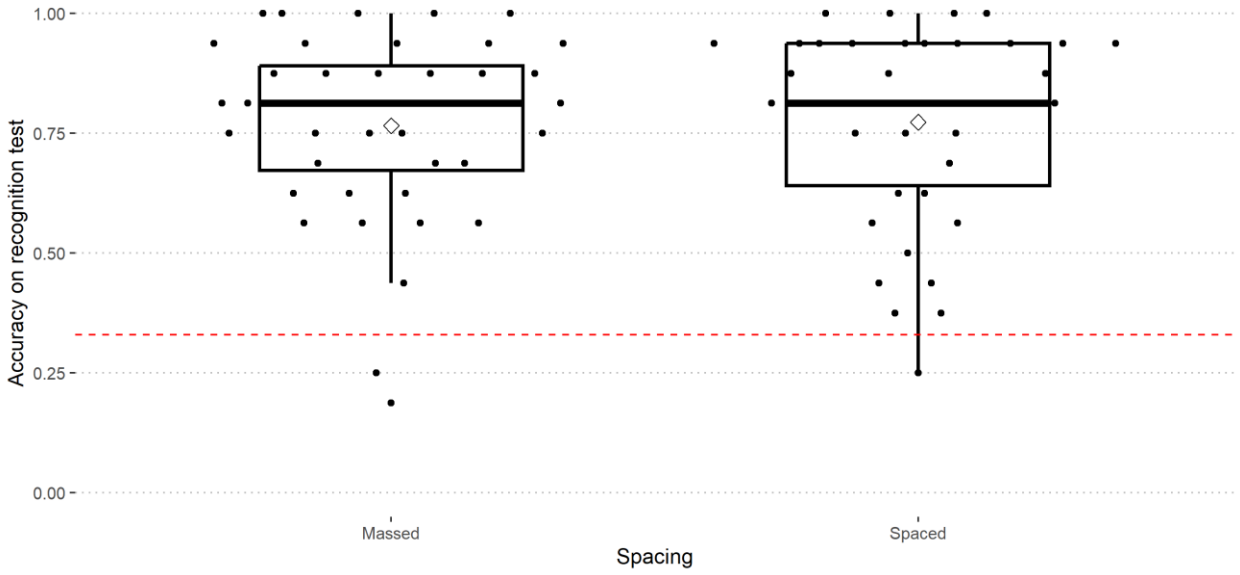
Note. A boxplot showing the percentage score on the cued recall task by spacing condition and complexity (linear versus relational category).

Accuracy on the recognition task was high (see Figure 4-9) and there were no significant differences in performance between the massed and spaced groups ($t = -0.163$, , d.f. = 75.11, $p = 0.870$, 95% CI = [-0.097, 0.082]). I also separated the types of mistakes participants could have made. Mistakes were either category based (i.e. they stated that a constraint from the linear category was from the relational category) or old/new mistakes (i.e. they responded that a constraint they had never seen before i.e. “Multiple of three?” was used in a category they did

learn) (see Figure 4-10). The types of mistakes participants made across spaced versus massed conditions were very similar and not insightful.

Figure 4-9

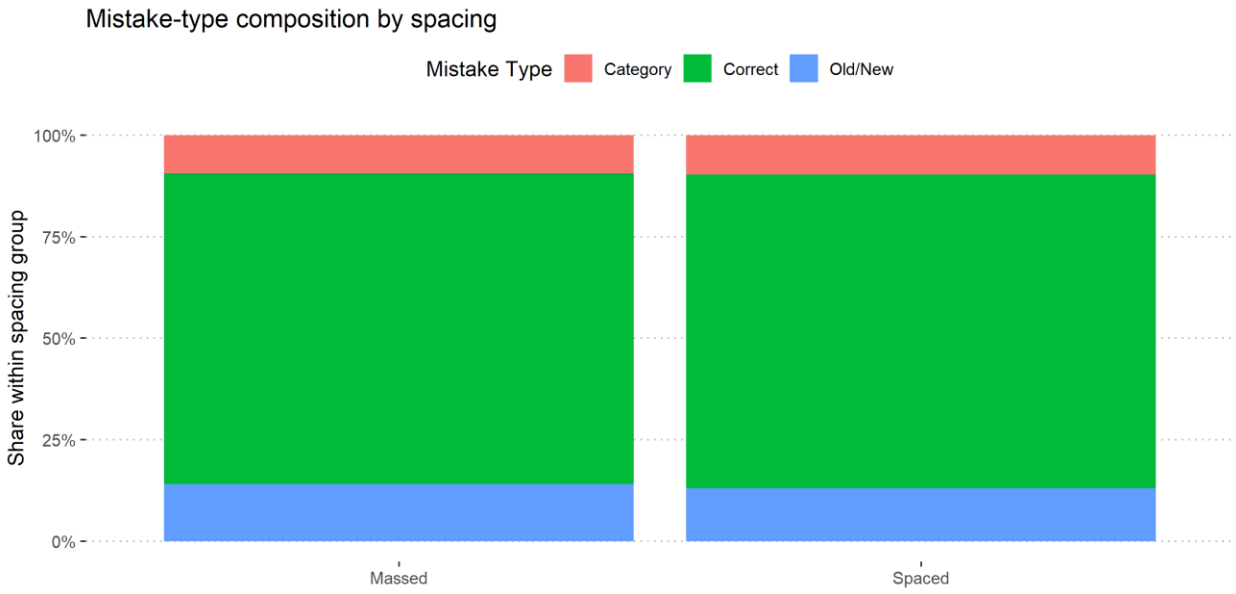
Experiment Three: Accuracy on the Recognition Task by Spacing Condition



Note. A boxplot showing the percentage score on the recognition test by spacing condition.

Figure 4-10

Experiment Three: Breakdown of Correct Answers versus the Types Of Mistakes Participants Made During Recognition Task by Spacing Condition



Note. A boxplot showing the percentage score on the recognition test by spacing condition.

5 General Discussion

I will first reiterate the aims of this thesis, then I will summarise the key findings from each chapter. Following this, I will discuss key themes that persisted throughout the thesis with a focus on retrieval during practice, individual differences, interference, and task complexity. Finally, I will provide an overall conclusion to the thesis.

The aim of this thesis was to investigate whether task complexity interacted with the spacing effect. I chose to investigate massed versus spaced practice as distributing practice is one of the top recommendations for educators to improve learners' retention over a longer period. It is frequently recommended by cognitive psychologists (Carpenter et al., 2012; Latimier et al., 2021), governments (Ofsted, 2021), and used in educational technology (Brown et al., 2021). For verbal facts and lists of words, there is considerable evidence for optimal spacing routines and accurate predictions of forgetting for remembering a series of isolated facts (Cepeda et al., 2008; Pavlik & Anderson, 2008; Walsh et al., 2022), however, none of these models account for differences in the complexity of the material. As a practical example, when a student is planning their revision should a new times table fact, such as "12 x 11 = 132", be practised again as soon as learning a new rule in calculus such as the chain rule? As current verbal theories and models of the spacing effect do not make clear predictions as to how task complexity may affect the efficacy of spacing, it is not currently possible to make such a recommendation.

Before a school can confidently, and purposefully, space practice across a curriculum, or an edtech platform can create an effective personalised study schedule for each student, four questions need to be answered:

1. Is spacing effective for complex material?
2. What aspects of the task or material increase complexity in a way which affects spacing?

3. If spacing is still effective, how does complexity affect the optimal spacing of material?
4. When spacing multiple learning objectives. How does spacing affect material that has links between the elements to be learned?

In this thesis I aimed to answer the first two questions. First, the experimental results alongside the meta-analysis provide strong evidence that spacing can be effective for learning mathematics, but more research needs to be done to show that testing is effective for mathematics relative to restudy. Across the four experiments presented in this thesis, I found no evidence that changes to procedural complexity or element interactivity interacted with the spacing effect.

5.1 Chapter 2 Summary

Chapter 2 presented a meta-analysis (published in Educational Psychology Review) assessing whether spaced practice, retrieval practice, or the combination of the two learning techniques are effective for learning mathematics. After running the initial search, it was apparent that there was insufficient evidence to address whether the combination of spacing and retrieval was effective for mathematics learning. While completing the meta-analysis I endeavoured to follow current best practices including pre-registration, I followed PRISMA guidelines (Page et al., 2021), used Covidence for reproducibility (Covidence Systematic Review Software, 2023), reported heterogeneity, and I accounted for the fact that most studies provided multiple effects using Robust Variance Estimation (Pustejovsky & Tipton, 2022). These steps helped to ensure that the results and conclusions were as valid as possible and should help with any future replications of the meta-analysis.

There was limited evidence for the efficacy of retrieval practice rather than restudying materials in mathematics learning. The search revealed seven studies for mathematics learning

that compared restudy versus testing and while there was a positive mean effect it was not robust, and the 95% confidence interval crossed zero. Given the substantial number of successful testing effect studies in other domains such as verbal learning, I did not conclude that this result suggested testing was not effective for mathematics learning. Instead, I suggest that this result highlights the need for future studies involving testing and mathematics to confirm that it is a robust effect. Another meta-analysis, with broader search parameters found 12 other papers not included in this study with different control conditions other than restudy (including testing versus no activity, testing versus elaborative strategies or more versus fewer test questions), suggesting that alternate research questions have been of greater interest to prior researchers (Yang et al., 2021). Perhaps prior researchers assumed the testing effect would hold for mathematics material and moved to questions of more or less testing or when testing occurs, before checking the robustness of the testing effect for mathematics material. My meta-analysis shows, however, that there is still a need for fundamental research into the testing effect and mathematics learning and it would be valuable to provide sufficient evidence for its adoption on a larger scale. For these future studies, the testing versus studying worked examples approach, as used by Yeo and Fazio (2019), would be a good approach to fill this gap.

Overall, spaced practice was effective for mathematics learning, with a small to medium effect size, however, the search revealed two distinct paradigms that researchers employ when conducting research into distributed practice and mathematics learning. Participants either learnt the required information in isolation or as part of a course alongside other material. When material was taught in isolation the effect was larger, possibly due to increased control under lab conditions, while the weighted mean effect was lower for material presented as part of a course, perhaps due to interference between items. Due to the sample size, the potential moderating

factors could not inform why isolated material benefits more from spaced practice or help to account for the large spread of effect sizes. I concluded that spacing was effective for mathematics learning.

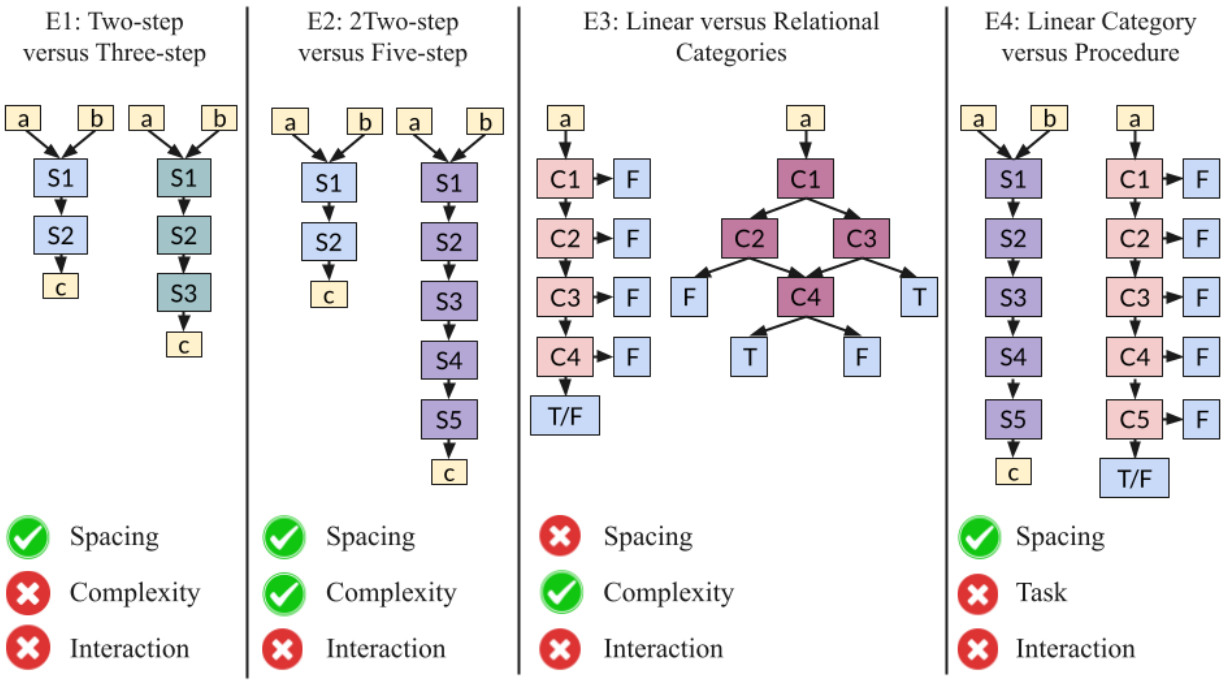
5.2 Experimental Work: Spacing Effect and Task Complexity

Chapter 3 and 4 describe the results of four experiments designed to investigate how the spacing effect changes when different aspects of task complexity are manipulated (see Figure 5-1). Across all experiments I employed the same spacing schedule. I compared massed practice, where participants completed all the practice trials in a single session, with spaced practice, where participants completed the practice trials distributed evenly over three consecutive days. There was then a final session, after a seven-day delay, during which participants completed a post-test. Previous studies had issues where participants in the spaced condition had greater attrition which could have introduced a bias into the study (Barzagar Nazari & Ebersbach, 2018). To minimise this potential bias all participants completed the four sessions and working memory, arithmetic fluency, and maths anxiety measures were organised to pad the days when massed participants did not complete the main task.

Across all experiments the main analysis was designed to detect a main effect of spacing (i.e., was there a significant difference between spaced and massed practice?), a main effect of Complexity/Task (i.e., was there a significant difference between the complexity or task conditions?) and finally whether there was a significant interaction between Spacing and complexity (i.e., was spacing more or less effective in one complexity condition or the other?).

Figure 5-1

Overview of Experimental Work



5.2.1 Chapter 3 Summary

Chapter 3 describes the results of two experiments where I varied the number of steps in mathematics procedures and compared a massed versus spaced practice schedule. These experiments were the first spacing studies to attempt to vary complexity within a mathematics learning task. I used procedures consisting of a series of arithmetic steps (i.e., "add the two numbers together" or "divide the answer by three"). The artificial nature of the procedures made the experiment less ecologically valid, but minimised the effect of prior knowledge, intending that the only change was the procedural complexity (the number of steps in the procedures).

I initially compared a two- versus three-step procedure in experiment one, before comparing a two- versus five-step procedure in experiment two. In experiment one, I found a

significant main effect of spacing, but I did not find a significant effect of complexity, nor a significant interaction effect between the two terms. This suggested that an additional step was not sufficient for a difference in complexity to be detected, so I added two extra steps to the complex condition for the next experiment. In experiment two, comparing a two- versus a five-step procedure, I found both a significant main effect of spacing and complexity, but no significant interaction effect between spacing and complexity. These results suggest that spacing is robust to changes in procedural complexity and another measure of complexity may be more important for educators to consider.

5.2.2 Chapter 4 Summary

Chapter 4 drew on research that was published during data collection of experiments one and two. Chen et al. (2023) published an article which provided an in-depth definition for element interactivity, which had previously been critiqued for not being sufficiently well operationalised (Karpicke & Aue, 2015). Furthermore, Chen et al. (2024) published a series of experiments exploring how element interactivity interacted with a version of the spacing effect they dubbed the resting effect. They found mixed results that showed a spacing effect when element interactivity was high (which they suggested was due to participants being able to recover working memory resources during rest) and when it was low (which they suggested was due to participants being able to rehearse the information during rest). In their third and fourth experiment, where they compared spaced and massed practice with high and low element interactivity, they used the same material and instead manipulated the expertise of participants. They had one expert group (low element interactivity) and a novice group (high element interactivity). I aimed to further this research in two ways. First, I wanted to find a task where I could manipulate element interactivity directly instead of manipulating participants' prior

knowledge. I did this by having participants learn two categories, a linear category, where each constraint could be learnt in isolation (low element interactivity) and a relational category, where each constraint was only useful when other elements were taken into consideration (high element interactivity). This was a useful next step as in Chen et al. (2024) 's experiment the expert and novice groups may have had a large variation in prior knowledge, while this material was designed so that each participant should have equal knowledge of the individual elements and only the structure of how elements are linked together was varied. Second, I used longer inter-session intervals and retrieval intervals. This is useful as shorter retrieval intervals can be a boundary condition of the spacing effect and at short intervals massed practice may be superior to spaced practice (Rohrer & Taylor, 2007).

I hypothesised that the linear category would benefit from spacing while the relational category would not because of the prior research showing that higher complexity leads to reduced efficacy of spacing (Donovan & Radosevich, 1999). While the working memory resource depletion hypothesis (Chen et al., 2024) would instead suggest that the relational category would benefit more from spacing as the intervals between the sessions would allow for working memory resources to be restored and lead to greater subsequent learning. In contrast to either hypothesis, neither the linear nor relational category benefited from spaced practice relative to massed practice. I suggested two reasons for this, either both categories were too high in element interactivity to benefit from spacing, in comparison to the procedures task used in chapter 3, or there were high levels of interference between the two category learning tasks that hindered learning.

In experiment four, participants were taught a five-step procedure (adapted from experiment two) and a five-constraint linear category (adapted from experiment three). This time

I found a significant effect of spacing across both task conditions, but not a main effect of task or an interaction between spacing and task (although it was the closest to significance of all the other experiments). This suggests that in experiment three the linear category in experiment one was not too high in element interactivity to benefit from the spacing effect, but instead it was the fact that it was learnt alongside the relational category. Following these two experiments I can make no strong statements about whether element interactivity interacts with the spacing effect. Furthermore, future experiments should avoid learning material where there is likely to be a lot of interference, manipulating the structure of the material while changing and counterbalancing the surface features of the task. Alternatively, researchers could explicitly manipulate the material so that there is a high interference condition and a low interference condition to see if this a potential boundary condition for the spacing effect in mathematics learning.

5.3 Theoretical and Methodological Considerations

After summarising the results of the meta-analysis and two experimental chapters I will now focus on four key themes that persisted across the thesis. First, I will discuss the role of retrieval during practice in mathematics learning and whether I saw the trade-off between lower early performance and greater long-term retention typical in spacing studies. Then, I will discuss the individual difference measures I tracked across the experiments. Next, I will expand on interference as the explanation for why I did not see a spacing effect in experiment four before a final discussion of task complexity and presenting overall conclusions.

Retrieval Accuracy on the Practice Trials. During data extraction of the meta-analysis, it was difficult to understand whether the included studies required participants to actively retrieve the material from memory or allowed them to use external aids, such as notes, examples, or prior work. As participants' retrieval of the material after initial acquisition is key to the study-phase

account of the spacing effect (Thios & D'Agostino, 1976), this seemed important and was one of the key ways the meta-analysis informed the design of the tasks used in chapters 2 and 3. Across all the experiments presented in this thesis, I wanted to ensure that after the initial learning phase participants would be required to retrieve the information. I asked participants not to write anything down and, after the initial learning session, participants could not see the procedure/category they were trying to learn on screen. However, participants were given feedback if incorrect, so after the first question participants may have seen the procedure/category again. If that was the case then they only had to rehearse the procedure/category in working memory, rather than recall from long-term memory, but this is still more explicit than many prior studies and something that should be included in mathematics learning and spacing research going forward. Furthermore, it would be difficult to rehearse the whole procedure or category at once. One way to test whether feedback reduces the efficacy of ensuring retrieval during practice could be to add a distractor between each practice trial so that participants cannot maintain the procedure or category in working memory, however this was not feasible in my experiments due to time constraints. I think that being explicit about what participants had to retrieve and when they were able to follow an example was an important methodological improvement.

Spacing did not routinely lead to poorer practice performance when compared to massed practice in three out of the four experiments presented in this thesis. Spacing practice has been described as a desirable difficulty, a learning technique that makes practice more difficult to the benefit of future learning and retention (Bjork & Bjork, 2020; Lyle et al., 2022). It is often a trade-off between poorer initial performance and greater future performance. In experiment four I found this to be the case, retrieval accuracy during practice was significantly higher in the

massed condition than in the spaced condition. This creates a potential danger when spacing practice in the classroom as students find it more difficult and may perform worse, reducing their confidence. In contrast, in experiments one, two, and three there was no significant difference in retrieval accuracy between spacing conditions. This is positive as participants benefited from spaced practice in experiments one and two while not suffering reduced performance during practice. However, as participants in the spaced condition in experiment four did find their practice more difficult, this highlights the importance of ensuring that students understand why they are being asked to space their practice, that it will be more difficult, but that it will help them to learn.

Individual differences. Across my four experimental studies I measured participants' working memory capacity (WMC), measured using forwards and backwards digit span tasks, arithmetic fluency, and mathematics anxiety (excluding experiment three where mathematics anxiety was omitted). For each experiment, there were no significant differences in working memory capacity or arithmetic fluency between the massed and spaced groups. This suggests that the differences in final performance were due to the spacing manipulation, not a bias introduced through our sampling of participants. For each experiment I ran exploratory analyses to detect whether any of the individual difference measures were significant covariates, in each case they were not significant. As these analyses were exploratory, they provided no potential insights, and I did not have sufficient power to detect small effects, I did not include them in the thesis, but the data and code to run them is available in the supplementary materials.

While it did not appear to have an effect in the studies presented in this thesis, there is some evidence to suggest that spacing may benefit participants with higher WMC more than those with lower WMC (Bui et al., 2013), however, other results suggest that participants with

higher WMC simply perform better on memory tests and higher WMC has an additive effect on performance and does not interact with the spacing effect (Delaney et al., 2018). Overall, while it is reasonable to expect participants with higher WMC to perform better on more complex material, I found no evidence that this is the case. If a future study finds a manipulation of task complexity that significantly affects the efficacy of spacing, then an additional high-powered study with multiple measures of working memory capacity would provide a better understanding of the relationship between WMC and the spacing effect.

In each experiment, I decided to test participants' mathematics anxiety at the end of the study, after the final test, so that taking the test itself did not alter participants' anxiety during the task. This presents a limitation as participation in any of the experiments could have either increased anxiety, if participants found the task difficult, or reduced it, perhaps due to increased fluency during massed practice leading to overconfidence (Emeny et al., 2021). I did not have the resources to test mathematics anxiety twice during the experiment, however, if mathematics anxiety is a focus of a future experiment, then an improvement would be to test participants at the start and end of the experiment. This would allow future researchers to better distinguish between a baseline mathematics anxiety of the participant and the change in mathematics anxiety during the experiment.

Due to a change in recruitment platform, participants were older in experiment four, than the first three experiments. In experiment one, two and three I recruited my participants using the University of York's human participant pool SONA and these participants were similar in age (around 20 years old). However, in experiment four I recruited participants via prolific, and the participants were older (around 27 years old). Before experiment four, our criteria were that participants needed to be over 18 and an undergraduate student and I used the same criteria on

prolific. Either prolific has a larger proportion of mature students, who signed up to our experiment quickly, or perhaps participants on prolific are incentivised to claim that they are a university student in order to gain access to more studies. However, I do not think that this caused a major issue as participants had high performance on the task and post-test, and previous research suggests the spacing effect is robust across many age ranges (Delaney et al., 2010; Kornell & Bjork, 2008; MacLean et al., 2017).

Interference. Additionally, further research should be undertaken to see how different items affect the spacing schedules of other items learnt alongside each other in practice. The meta-analysis presented in chapter 2 suggested that a smaller spacing effect was found for items presented together in a course and the material in experiment three did not benefit from spacing when presented together, while an adapted version of the same item in experiment four did benefit when learnt alongside less similar material. This is not a ubiquitous finding, however, as the two procedures used in experiments one and two were visually and conceptually similar and they consistently benefited from spaced practice relative to massed practice. Future studies that manipulate the structure of categories, but vary the surface features, would be a valuable next step. As an example, one task could require participants to categorise a number, while the other requires participants to categorise a shape. Another future experiment could include a direct manipulation of spaced versus massed practice, high versus low element interactivity categories and similar versus dissimilar surface features. This 2 x 2 x 2 factorial design could test whether the high levels of interference across tasks may be a boundary condition for the spacing effect.

Task Complexity. The most difficult part of research into task complexity is that it is unique to each participant. There is evidence that spacing benefits “complex” material, such as learning calculus (Hopkins et al., 2016; Lyle et al., 2020, 2022) and some people may argue that spacing

has already been shown for complex material. After all, the brain has a remarkable ability to chunk information together allowing us to gain expertise and make the complex simple (Gobet et al., 2001; Miller, 1956). However, if participants had high prerequisite prior knowledge and had portions of the "complex" material chunked into a single element then that material was not complex to that participant. Across four experiments I aimed to manipulate complexity and practice schedule while reducing the effect of prior knowledge. My key contribution was the use of artificial elements (i.e. a step to perform "multiply by five" or a constraint to check "is the number a multiple of five?") to build up novel learning tasks in a modular fashion. Importantly, I expected that each individual element would be known to the participants (and checked this during piloting) and that what they are learning is the structure of the procedure or category. By varying the structure, I sought to manipulate first procedural complexity in experiments 1 and 2, and then element interactivity in experiments 3 and 4.

When measuring procedural complexity as the number of steps participants were required to learn, we found large main effects of spacing. Further evidence, with larger sample sizes, would be required to make a strong recommendation to educators that it can be ignored when scheduling practice, however in my experiments it did not affect spacing. My manipulation of element interactivity found no effects of spacing, but this was likely to not be due to the material itself, but instead to how the task was set up. The linear category learning task did not benefit from spacing when learnt alongside a similar relational category learning task. However, the linear category did benefit from spacing when paired with a procedure learning task. If element interactivity does matter, then one methodological concern future experiments should consider is to avoid interference between learning objectives. Overall, the experiments add to the consensus that the spacing effect is robust across different tasks.

One explanation for the spacing effects robustness to changes in task complexity could be due to multiple complimentary mechanisms underlying the spacing effect. Most modern overviews of the spacing effect argue that no single mechanism that can account for, and model, all the key phenomena of the spacing effect (Delaney et al., 2010; Küpper-Tetzel, 2014; Walsh et al., 2018). For example, if more complex material with more elements increases the difficulty of retrieval then the study-phase retrieval account would suggest a reduced effectiveness of spacing (Thios & D'Agostino, 1976), however perhaps if participants realise this during practice, either consciously or subconsciously, they may attend to the information more closely. If this occurs, then in the massed condition participants may feel fluent in the material and process it less deeply, as suggested in the deficient processing account (Hintzman, 1974), while they may attend more closely to the complex material in the spaced condition when unable to recall the material.

I was unable to test this theory of multiple complementary mechanisms in the experiments presented in this thesis; however, future researchers could test this theory by running a series of experiments manipulating inter-session intervals and amount of practice trials. One study found that increasing the amount of practice beyond the minimum needed was found not to moderate the spacing effect (Lyle et al., 2020; Rohrer & Taylor, 2006), however, this may have been due to the use of a simple learning objective. Increasing or reducing the number of practice sessions can enable a researcher to see whether deficient processing is occurring as at some point the less complex material will begin to no longer benefit from increased practice while the more complex may benefit from additional practice. Additionally, if the critical point where deficient processing occurs is moderated by changes to the inter-session interval then that may be evidence that participants are using the difficulty of retrieval as a sign

to process the material at a deeper level. If this is indeed the case, then this would make spacing an even more valuable tool to improve learning.

5.4 Overall Conclusions

I found no evidence that procedural complexity or element interactivity affects the efficacy of spaced practice. Importantly, however, my manipulation of element interactivity was impeded by high interference between the two tasks and should not be considered evidence that element interactivity does not affect the efficacy of spacing. More evidence is needed to show procedural complexity has no effect on the efficacy of spacing, however, future research should focus on explicit manipulations of element interactivity that use different surface features, to reduce interference, while manipulating the structure of the materials. If spacing is affected by task complexity, then further studies should vary the inter-session and retrieval intervals to ascertain how task complexity affects the optimal spacing schedules for different tasks. Taken together, the results of the meta-analysis and robustness of the spacing effect across three out of four experiments presented in this thesis supports the inclusion of spaced practice in mathematics learning.

6 References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Rev. Educ. Res.*, *87*(3), 659–701. ***https://doi.org/10.3102/0034654316689306***

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. ***https://doi.org/10.3758/s13428-019-01237-x***

ARC Education. (n.d.). *Why it works - ARC education*. ***https://www.arceducation.co.uk/why-it-works/***

Balota, D. A., Duchek, J. M., & Paullin, R. (1989). Age-related differences in the impact of spacing, lag, and retention interval. *Psychology and Aging*, *4*(1), 3–9. ***https://doi.org/10.1037/0882-7974.4.1.3***

Barzagar Nazari, K., & Ebersbach, M. (2018). Distributed practice: Rarely realized in self-regulated mathematical learning. *Frontiers in Psychology*, *9*(NOV), 2170. ***https://doi.org/10.3389/fpsyg.2018.02170***

Barzagar Nazari, K., & Ebersbach, M. (2019). Distributed practice in mathematics: Recommendable especially for students on a medium performance level? *Trends in Neuroscience and Education*, *17*, 100122. ***https://doi.org/10.1016/j.tine.2019.100122***

Beagley, J. E., & Capaldi, M. (2016). The effect of cumulative tests on the final exam. *PRIMUS: Problems, Resources, and Issues in Mathematics Undergraduate Studies*, *26*(9), 878–888. ***https://doi.org/10.1080/10511970.2016.1194343***

Beagley, J. E., & Capaldi, M. (2020). Using cumulative homework in calculus classes. *PRIMUS: Problems, Resources, and Issues in Mathematics Undergraduate Studies*, *30*(3), 335–348. ***https://doi.org/10.1080/10511970.2019.1588814***

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology*, *41*(2), 257–278. ***https://doi.org/10.1111/j.2044-8317.1988.tb00901.x***

Bego, C. R., Lyle, K. B., Ralston, P. A., & Hieb, J. L. (2017). *2017 IEEE frontiers in education conference (FIE)*. *2017-October* 1–5. ***https://doi.org/10.1109/FIE.2017.8190463***

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? *Cogn. Psychol.*, *61*(3), 228–247. ***https://doi.org/10.1016/j.cogpsych.2010.05.004***

Bercovitz, K. E., Bell, M. C., Simone, P. M., & Wiseheart, M. (2017). The spacing effect in older and younger adults: does context matter? *Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition*, *24*(6), 703–716. ***https://doi.org/10.1080/13825585.2016.1251552***

Betsch, T., Quittenbaum, N., & Lüders, M. (2015). On the robustness of the quizzing effect under real teaching conditions. *Z. Entwicklungspsychol. Padagog. Psychol.*, *29*(2), 109–114. ***https://doi.org/10.1024/1010-0652/a000149***

Bjork, E. L., Bjork, R., Roediger, H. L., Mcdermott, K. B., & Mcdaniel, M. A. (2011). *Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning*. ***http://mrbartonmaths.com/resourcesnew/8.%20Research/Memory%20and%20Revision/Making-Things-Hard-on-Yourself-but-in-a-Good-Way-2011.pdf***

Bjork, R. A., & Bjork, E. L. (2020). Desirable difficulties in theory and practice. *Journal of Applied Research in Memory and Cognition*, *9*(4), 475–479. ***https://doi.org/10.1016/j.jarmac.2020.09.003***

Blaxton, T. A. (1989). Investigating dissociations among memory measures: Support for a transfer-appropriate processing framework. *J. Exp. Psychol. Learn. Mem. Cogn.*, *15*(4), 657–668. ***https://doi.org/10.1037/0278-7393.15.4.657***

Brown, E. R., Culora, A., & Ilie, S. (2021). *Independent analysis of the relationship between Sparx Maths and maths outcomes*. ***https://doi.org/10.1007/978-3-658-28670-5_16***

Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin*, *145*(11), 1029–1052. ***https://doi.org/10.1037/bul0000209***

Bui, D. C., Maddox, G. B., & Balota, D. A. (2013). The roles of working memory and intervening task difficulty in determining the benefits of repetition. *Psychonomic Bulletin & Review*, *20*(2), 341–347. ***https://doi.org/10.3758/s13423-012-0352-5***

Bullard, J. C. (2020). Evaluating explicit timing modifications to maximize student learning rates: A comparison of distributed practice and goal setting with performance feedback. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, *81*(4-A), No–Specified. ***http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc17&NEWS=N&AN=2020-28119-064***

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. ***https://doi.org/10.1038/nrn3475***

Camp, J. S. (1973). The effects of distributed practice upon learning and retention in introductory algebra. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, *7328190*(5-A), 2455–2456. ***http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc2&NEWS=N&AN=1976-03010-001***

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *J. Exp. Psychol. Learn. Mem. Cogn.*, *35*(6), 1563–1569. ***https://doi.org/10.1037/a0017021***

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, *34*, 268-276.

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*(3), 369–378. ***https://doi.org/10.1007/s10648-012-9205-z***

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychol. Bull.*, *132*(3), 354–380. ***https://doi.org/10.1037/0033-2909.132.3.354***

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention: Research article. *Psychol. Sci.*, *19*(11), 1095–1102. ***https://doi.org/10.1111/J.1467-9280.2008.02209.X***

Chen, O., & Kalyuga, S. (2019). Cognitive load theory, spacing effect, and working memory resources depletion. In *Form, function, and style in instructional design: Emerging research and opportunities* (pp. 1–26). ***https://doi.org/10.4018/978-1-5225-9833-6.ch001***

Chen, O., & Kalyuga, S. (2021). Working memory resources depletion makes delayed testing beneficial. *J. Cogn. Educ. Psychol.* ***https://connect.springerpub.com/content/sgrjcep/20/1/38.abstract***

Chen, O., Castro-Alonso, J. C., Paas, F. & Sweller, J.; (2018). Extending cognitive load theory to incorporate working memory resource depletion: Evidence from the spacing effect. *Educ. Psychol. Rev.*, *30*(2), 483–501. ***https://doi.org/10.1007/s10648-017-9426-2***

Chen, O., Kai Yin Chan, B., Anderson, E., O'sullivan, R., Jay, T., Ouwehand, K., Paas, F., & Sweller, J. (2024). The effect of element interactivity and mental rehearsal on working memory resource depletion and the spacing effect. *Contemporary Educational Psychology*, *77*, 102281. ***https://doi.org/10.1016/j.cedpsych.2024.102281***

Chen, O., Paas, F., & Sweller, J. (2023). A cognitive load theory approach to defining and measuring task complexity through element interactivity. *Educational Psychology Review*, *35*(2), 63. ***https://doi.org/10.1007/s10648-023-09782-w***

Cheung, M. W. L. (2014). Modeling dependent effect sizes with three-level meta-analyses: a structural equation modeling approach. *Psychological Methods*, *19*(2), 211–229. ***https://doi.org/10.1037/a0032968***

Chung, Y., Rabe-Hesketh, S., & Choi, I.-H. (2013). Avoiding zero between-study variance estimates in random-effects meta-analysis. *Statistics in Medicine*, *32*(23), 4071–4089. ***https://doi.org/10.1002/sim.5821***

Covidence systematic review software. (2023). Veritas Health Innovation. ***www.covidence.org.***

Crissinger, B. R. (2015). The effect of distributed practice in undergraduate statistics homework sets: A randomized trial. *J. Stat. Educ.*, *23*(3). ***https://doi.org/10.1080/10691898.2015.11889743***

Crooks, N. M., & Alibali, M. W. (2014). Defining and measuring conceptual knowledge in mathematics. *Dev. Rev.*, *34*(4), 344–377. ***https://doi.org/10.1016/j.dr.2014.10.001***

Delaney, P. F., Godbole, N. R., Holden, L. R., & Chang, Y. (2018). Working memory capacity and the spacing effect in cued recall. *Memory*, *26*(6), 784–797. ***https://doi.org/10.1080/09658211.2017.1408841***

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* (Vol. 53, pp. 63–147). ***https://doi.org/10.1016/S0079-7421(10)53003-2***

Dempster. (1988). *The spacing effect a case study in the failure to apply the results of psychological research*.

Dirkx, K. J. H., Kester, L., & Kirschner, P. A. (2014). The testing effect for learning principles and procedures from texts. *J. Educ. Res.*, *107*(5), 357–364. ***https://doi.org/10.1080/00220671.2013.823370***

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology*, *84*(5), 795. ***https://doi.org/10.1037/0021-9010.84.5.795***

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463. ***https://doi.org/10.1111/j.0006-341x.2000.00455.x***

Ebbinghaus, H. (1913). Memory: A contribution to experimental psychology. (H. A. Ruger & C. E. Bussenius, Trans.). Teachers College Press. https://doi.org/10.1037/10011-000

Ebersbach, M., & Barzagar Nazari, K. (2020a). Implementing distributed practice in statistics courses: Benefits for retention and transfer. *J. Appl. Res. Mem. Cogn.*, *9*(4), 532–541. ***https://doi.org/10.1016/j.jarmac.2020.08.014***

Ebersbach, M., & Barzagar Nazari, K. (2020b). No robust effect of distributed practice on the short- and long-term retention of mathematical procedures. *Frontiers in Psychology*, *11*, 811. ***https://doi.org/10.3389/fpsyg.2020.00811***

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), 629–634. ***https://doi.org/10.1136/bmj.315.7109.629***

Emeny, W. G., Hartwig, M. K., & Rohrer, D. (2021). Spaced mathematics practice improves test scores and reduces overconfidence. *Applied Cognitive Psychology*, *35*(4), 1082–1089. ***https://doi.org/10.1002/acp.3814***

Eustace, J., Bradford, M., & Pathak, P. (2020). Evaluating flipped and traditional pedagogy with retrieval practice in mathematics for computing. *Journal of Computers in Mathematics and Science Teaching*, *39*(2), 149–167. ***https://www.learntechlib.org/primary/p/210971/***

Fazio, L. K. (2018). *The effects of retrieval practice on fraction arithmetic knowledge* (pp. 169–182). Routledge/Taylor & Francis Group. ***https://doi.org/10.4324/9781315200446-11***

Foster, N. L., Mueller, M. L., Was, C., Rawson, K. A., & Dunlosky, J. (2019). Why does interleaving improve math learning? The contributions of discriminative contrast and distributed practice. *Mem. Cognit.*, *47*(6), 1088–1101. ***https://doi.org/10.3758/s13421-019-00918-4***

Gavaghan, D. J., Moore, R. A., & McQuay, H. J. (2000). An evaluation of homogeneity tests in meta-analyses in pain using simulations of individual patient data. *Pain*, *85*(3), 415–424. ***https://doi.org/10.1016/S0304-3959(99)00302-4***

Gay, L. R. (1973). Temporal position of reviews and its effect on the retention of mathematical rules. *Journal of Educational Psychology*, *64*(2), 171–182. ***https://doi.org/10.1037/h0034595***

Gidaropoulos, A. (2021). *Hegarty maths: Gimmick or game changer?* [PhD thesis]. ***https://www.proquest.com/docview/2700375371/abstract/2CDB8768A5274D93PQ/1***

Glenberg, A. M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Mem. Cognit.*, *7*(2), 95–112. ***https://doi.org/10.3758/bf03197590***

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C.-H., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, *5*(6), 236–243. ***https://doi.org/10.1016/s1364-6613(00)01662-4***

Goettl, B. P., Yadrick, R. M., Connolly-Gomez, C., Regian, J. W., & Shebilske, W. L. (1996). Alternating task modules in isochronal distributed training of complex tasks. *Hum. Factors*, *38*(2), 330–346. ***https://doi.org/10.1518/001872096779048048***

Gorgievski, N. (2012). The impact of the spacing effect and overlearning on student performance in calculus. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, *3464319*(9-A), 3178. ***http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc9&NEWS=N&AN=2012-99050-453***

Greene, R. L. (1989). Spacing effects in memory: Evidence for a two-process account. *J. Exp. Psychol. Learn. Mem. Cogn.*, *15*(3), 371–377. ***https://doi.org/10.1037/0278-7393.15.3.371***

Hake, S. (2012). *Saxon math. Course 1*. Houghton Mifflin Harcourt Publishing Company.

Hansen, C., Steinmetz, H., & Block, J. (2022). How to conduct a meta-analysis in eight steps: A practical guide. *Management Review Quarterly*, *72*(1). ***https://doi.org/10.1007/s11301-021-00247-4***

Harrer, M., Cuijpers, P., Furukawa, T., & Ebert, D. D. (2019). *Dmetar: Companion R package for the guide 'doing meta-analysis in R'* [Manual].

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hedges, L. V. (1981). Distribution Theory for Glass's Estimator of Effect size and Related Estimators. *Journal of Educational Statistics*, *6*(2), 107–128. ***https://doi.org/10.3102/10769986006002107***

Hiebert, J., & Lefevre, P. (1986). Procedural and conceptual knowledge. *Conceptual and Procedural Knowledge: The Case of Mathematics*, 127.

Hintzman, D. L. (1974). Theoretical implications of the spacing effect. *Theories in Cognitive Psychology: The Loyola Symposium.*, *386*. ***https://psycnet.apa.org/fulltext/1975-00291-001.pdf***

Hintzman, D. L. (2004). Judgment of frequency versus recognition confidence: Repetition and recursive reminding. *Mem. Cognit.*, *32*(2), 336–350. ***https://doi.org/10.3758/bf03196863***

Hintzman, D. L., Block, R. A., & Summers, J. J. (1973). Modality tags and memory for repetitions: Locus of the spacing effect. *J. Verbal Learning Verbal Behav.*, *12*(2), 229–238. ***https://doi.org/10.1016/s0022-5371(73)80013-1***

Hirsch, C. R., Kapoor, S. F., & Laing, R. A. (1982). Alternative models for mathematics assignments. *Internat. J. Math. Ed. Sci. Tech.*, *13*(3), 243–252. ***https://doi.org/10.1080/0020739820130301***

Hjetland, H. N., Brinchmann, E. I., Scherer, R., Hulme, C., & Melby-Lervåg, M. (2020). Preschool pathways to reading comprehension: A systematic meta-analytic review. *Educational Research Review*, *30*, 100323. ***https://doi.org/10.1016/j.edurev.2020.100323***

Holdan, E. G. (1986). A comparison of the effects of traditional, exploratory, distributed, and a combination of distributed and exploratory practice on initial learning, transfer, and retention of verbal problem types in first-year algebra. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, *46*(9-A), 2542. ***http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc2&NEWS=N&AN=1986-55259-001***

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***

Hunt, T. E., Clark-Carter, D., & Sheffield, D. (2011). The development and part validation of a U. K. Scale for mathematics anxiety. *Journal of Psychoeducational Assessment*, *29*(5), 455–466. ***https://doi.org/10.1177/0734282910392892***

Karpicke, J. D. (2012). Retrieval-based learning: Active retrieval promotes meaningful learning. *Curr. Dir. Psychol. Sci.*, *21*(3), 157–163. ***https://doi.org/10.1177/0963721412443552***

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educ. Psychol. Rev.*, *27*(2), 317–326. ***https://doi.org/10.1007/s10648-015-9309-3***

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). *Retrieval-based learning: An episodic context account* (B. H. Ross, Ed.; Vol. 61, pp. 237–284). Elsevier Academic Press, x. ***https://psycnet.apa.org/fulltext/2014-12777-007.pdf***

Kim, S. K., & Webb, S. (2022). The Effects of Spaced Practice on Second Language Learning: A Meta-Analysis. *Language Learning*, *72*(1), 269–319. ***https://doi.org/10.1111/lang.12479***

Kornell, N., & Bjork, R. A. (2008). Learning Concepts and Categories: Is Spacing the "Enemy of Induction"? *Psychological Science*, *19*(6), 585–592. ***https://doi.org/10.1111/j.1467-9280.2008.02127.x***

Kornell, N., Castel, A. D., Eich, T. S., & Bjork, R. A. (2010). Spacing as the friend of both memory and induction in young and older adults. *Psychology and Aging*, *25*(2), 498–503. ***https://doi.org/10.1037/a0017807***

Koval, N. G. (2019). Testing the deficient processing account of the spacing effect in second language vocabulary learning: Evidence from eye tracking. *Applied Psycholinguistics*, *40*(5), 1103–1139. ***https://doi.org/10.1017/S0142716419000158***

Küpper-Tetzel, C. E. (2014). Understanding the distributed practice effect. *Zeitschrift Für Psychologie*, *222*(2), 71–81. ***https://doi.org/10.1027/2151-2604/a000168***

Lakens, D., & Caldwell, A. R. (2021). Simulation-based power analysis for factorial analysis of variance designs. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920951503. ***https://doi.org/10.1177/2515245920951503***

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***

Lerma, A. (1990). *The effects of a cumulative review homework assignment schedule upon achievement in intermediate algebra and attitude toward mathematics* [PhD thesis].

https://www.proquest.com/dissertations-theses/effects-cumulative-review-homework-assignment/docview/303918610/se-2

Loenneker, H., Cipora, K., Artemenko, C., Nuerk, H.-C., & Huber, J. (2022). *Introducing the Math4Speed - a normed speeded test of arithmetic fluency*. ***https://osf.io/8mtpe/download***

Logan, J. M., Castel, A. D., Haber, S., & Viehman, E. J. (2012). Metacognition and the spacing effect: the role of repetition, feedback, and instruction on judgments of learning for massed and spaced rehearsal. *Metacognition and Learning*, *7*(3), 175–195. ***https://doi.org/10.1007/s11409-012-9090-3***

Lohnas, L. J., Polyn, S. M., & Kahana, M. J. (2011). Contextual variability in free recall. *J. Mem. Lang.*, *64*(3), 249–255. ***https://doi.org/10.1016/j.jml.2010.11.003***

Lyle, K. B., & Crawford, N. A. (2011). Retrieving essential material at the end of lectures improves performance on statistics exams. *Teach. Psychol.*, *38*(2), 94–97. ***https://doi.org/10.1177/0098628311401587***

Lyle, K. B., Bego, C. R., Hopkins, R. F., Hieb, J. L., & Ralston, P. A. S. (2020). How the amount and spacing of retrieval practice affect the short- and long-term retention of mathematics knowledge. *Educational Psychology Review*, *32*(1), 277–295. ***https://doi.org/10.1007/s10648-019-09489-x***

Lyle, K. B., Bego, C. R., Ralston, P. A. S., & Immekus, J. C. (2022). Spaced retrieval practice imposes desirable difficulty in calculus learning. *Educ. Psychol. Rev.*, *34*(3), 1799–1812. ***https://doi.org/10.1007/s10648-022-09677-2***

Onstwedder, E. (2023). *Duolingo*. ***https://blog.duolingo.com/spaced-repetition-for-learning/***

MacLean, A. C., Bell, M. C., & Simone, P. M. (2017). Interference Effects, Age, and the Spacing Benefit. *The American Journal of Psychology*, *130*(3), 295–302. ***https://doi.org/10.5406/amerjpsyc.130.3.0295***

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***

Magliero, A. (1983). Pupil dilations following pairs of identical and related to-be-remembered words. *Mem. Cognit.*, *11*(6), 609–615. ***https://doi.org/10.3758/bf03198285***

Mair, P., & Wilcox, R. (2020). Robust statistical methods in r using the WRS2 package. *Behavior Research Methods*, *52*. ***https://doi.org/10.3758/s13428-019-01246-w***

Mattis, K. V. (2015). Flipped classroom versus traditional textbook instruction: Assessing accuracy and mental effort at different levels of mathematical complexity. *Technology, Knowledge and Learning*, *20*(2), 231–248. ***https://doi.org/10.1007/s10758-014-9238-0***

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97. ***https://doi.org/10.1037/h0043158***

Morey, R. D., & Rouder, J. N. (2023). *BayesFactor: Computation of bayes factors for common designs*. ***https://CRAN.R-project.org/package=BayesFactor***

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519-533.

Moss, V. D. (1996). The efficacy of massed versus distributed practice as a function of desired learning outcomes and grade level of the student. *Dissertation Abstracts International:*

Section B: The Sciences and Engineering, *9603493*(9-B), 5204. ***http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc3&NEWS=N&AN=1996-95005-375***

Murre, J. M. J., & Dros, J. (2015). Replication and analysis of Ebbinghaus' forgetting curve. *PLoS One*, *10*(7), e0120644. ***https://doi.org/10.1371/journal.pone.0120644***

Ofsted. (2021). *Research review series: mathematics*. ***https://www.gov.uk/government/publications/research-review-series-mathematics***

Page, M. J., Moher, D., Fidler, F. M., Higgins, J. P. T., Brennan, S. E., Haddaway, N. R., Hamilton, D. G., Kanukula, R., Karunananthan, S., Maxwell, L. J., McDonald, S., Nakagawa, S., Nunan, D., Tugwell, P., Welch, V. A., & McKenzie, J. E. (2021). The REPRISE project: Protocol for an evaluation of REProducibility and replicability in syntheses of evidence. *Syst. Rev.*, *10*(1). ***https://doi.org/10.1186/s13643-021-01670-0***

Pan, S. C., & Carpenter, S. K. (2023). Prequestioning and pretesting effects: A review of empirical research, theoretical perspectives, and implications for educational practice. *Educ. Psychol. Rev.*, *35*(4), 97. ***https://doi.org/10.1007/s10648-023-09814-5***

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, *144*(7), 710–756. ***https://doi.org/10.1037/bul0000151***

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

Pavlik, P. I., Jr, & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect., *29*(4), 559–586. ***https://doi.org/10.1207/s15516709cog0000_14***

Powell, S. L. (2022). A comparative analysis of math facts fluency gains made through massed and distributed practice with varied inter-session intervals. *Dissertation Abstracts International: Section B: The Sciences and Engineering*, *28150257*(3-B), No–Specified. ***http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc18&NEWS=N&AN=2021-94593-063***

Pustejovsky, J. E., & Tipton, E. (2022). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prev. Sci.*, *23*(3), 425–438. ***https://doi.org/10.1007/s11121-021-01246-3***

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

Raaijmakers, J. (2003). Spacing and repetition effects in human memory: Application of the SAM model. *Cognitive Science, 27*(3), 431–452. ***https://doi.org/10.1016/s0364-0213(03)00007-7***

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. ***http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc2&NEWS=N&AN=1986-07610-001***

Reed, B. W. (1983). *Incremental, continuous-review versus conventional teaching of algebra*. search.proquest.com. ***https://search.proquest.com/openview/b35a0125a60bc0d0ad5733df1e9dae4b/1?pq-origsite=gscholar&cbl=18750&diss=y***

Reed, H. B. (1924). Distributed practice in addition. *J. Educ. Psychol.*, *15*(4), 248–249. ***https://doi.org/10.1037/h0070683***

Reynolds, M. R., Niileksela, C. R., Gignac, G. E., & Sevillano, C. N. (2022). Working memory capacity development through childhood: A longitudinal analysis. *Developmental Psychology*, *58*(7), 1254–1263. ***https://doi.org/10.1037/dev0001360***

Rickard, T. C., & Pan, S. C. (2018). A dual memory theory of the testing effect. *Psychonomic Bulletin & Review*, *25*(3), 847–869. ***https://doi.org/10.3758/s13423-017-1298-4***

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

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychol. Sci.*, *17*(3), 249–255. ***https://doi.org/10.1111/j.1467-9280.2006.01693.x***

Roediger, H. L., & Pyc, M. A. (2012). Applying cognitive psychology to education: Complexities and prospects. *Journal of Applied Research in Memory and Cognition, 1*(4), 263–265. ***https://doi.org/10.1016/j.jarmac.2012.10.006***

Rohrer, D., & Pashler, H. (2007). Increasing retention without increasing study time. *Curr. Dir. Psychol. Sci.*, *16*(4), 183–186. ***https://doi.org/10.1111/j.1467-8721.2007.00500.x***

Rohrer, D., & Taylor, K. (2006). The effects of overlearning and distributed practise on the retention of mathematics knowledge. *Applied Cognitive Psychology*, *20*(9), 1209–1224. ***https://doi.org/10.1002/acp.1266***

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning | instructional science. *Instructional Science*, *35*(6), 481–489. ***https://link.springer.com/article/10.1007/s11251-007-9015-8***

Rohrer, D., Dedrick, R. F., & Hartwig, M. K. (2020). The scarcity of interleaved practice in mathematics textbooks. *Educ. Psychol. Rev.*, *32*(3), 873–883. ***https://doi.org/10.1007/s10648-020-09516-2***

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. ***https://doi.org/10.1037/0033-2909.86.3.638***

Ross, B. H., & Landauer, T. K. (1978). Memory for at least one of two items: Test and failure of several theories of spacing effects. *J. Verbal Learning Verbal Behav.*, *17*(6), 669–680. ***https://doi.org/10.1016/s0022-5371(78)90403-6***

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

Satterthwaite, F. E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin*, *2*(6), 110–114. ***https://doi.org/10.2307/3002019***

Schutte, G. M., Duhon, G. J., Solomon, B. G., Poncy, B. C., Moore, K., & Story, B. (2015). A comparative analysis of massed vs. Distributed practice on basic math fact fluency growth rates. *Journal of School Psychology*, *53*(2), 149–159. ***https://doi.org/10.1016/j.jsp.2014.12.003***

Shaughnessy, J. J., Zimmerman, J., & Underwood, B. J. (1972). Further evidence on the MP-DP effect in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, *11*(1), 1–12. ***https://doi.org/10.1016/S0022-5371(72)80053-7***

Shea, J. B., & Morgan, R. L. (1979). Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(2), 179–187. ***https://doi.org/10.1037/0278-7393.5.2.179***

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***

Sommet, N., Weissman, D. L., Cheutin, N., & Elliot, A. J. (2023). How Many Participants Do I Need to Test an Interaction? Conducting an Appropriate Power Analysis and Achieving Sufficient Power to Detect an Interaction. *Advances in Methods and Practices in Psychological Science*, *6*(3), 25152459231178728. ***https://doi.org/10.1177/25152459231178728***

Sparx maths. (n.d.). ***https://sparxmaths.com/***

Steel, P., Beugelsdijk, S., & Aguinis, H. (2021). The anatomy of an award-winning meta-analysis: Recommendations for authors, reviewers, and readers of meta-analytic reviews. *Journal of International Business Studies*, *52*(1), 23–44. ***https://doi.org/10.1057/s41267-020-00385-z***

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. ***https://doi.org/10.1016/0364-0213(88)90023-7***

Sweller, J. (2006). The worked example effect and human cognition. *Learning and Instruction*, *16*(2), 165–169. ***https://doi.org/10.1016/j.learninstruc.2006.02.005***

Thios, S. J., & D'Agostino, P. R. (1976). Effects of repetition as a function of study-phase retrieval. *Journal of Verbal Learning & Verbal Behavior*, *15*(5), 529–536. ***https://doi.org/10.1016/0022-5371(76)90047-5***

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychol. Methods*, *20*(3), 375–393. ***https://doi.org/10.1037/met0000011***

Toppino, T. C., & Gerbier, E. (2014). *About practice: Repetition, spacing, and abstraction* (Vol. 60, pp. 113–189). Elsevier. ***https://www.sciencedirect.com/science/article/pii/B9780128000908000044***

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. ***https://doi.org/10.18637/jss.v036.i03***

Viechtbauer, W. (2017). *[R-meta] confidence interval for i^2 in multilevel model*. ***https://stat.ethz.ch/pipermail/r-sig-meta-analysis/2017-August/000150.html***

Viechtbauer, W., & Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, *1*(2), 112–125. ***https://doi.org/10.1002/jrsm.11***

Vincent, J., & Stacey, K. (2008). Do mathematics textbooks cultivate shallow teaching? Applying the TIMSS Video Study criteria to Australian eighth-grade mathematics textbooks. *Mathematics Education Research Journal*, *20*(1), 82–107. ***https://doi.org/10.1007/BF03217470***

Wahlheim, C. N., Maddox, G. B., & Jacoby, L. L. (2014). The role of reminding in the effects of spaced repetitions on cued recall: Sufficient but not necessary. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(1), 94–105. ***https://doi.org/10.1037/a0034055***

Walsh, M. M., Gluck, K. A., Gunzelmann, G., Jastrzembski, T., & Krusmark, M. (2018). Evaluating the theoretic adequacy and applied potential of computational models of the spacing effect. *Cognitive Science*, *42 (3)*, 644–691. ***https://doi.org/10.1111/cogs.12602***

Walsh, M. M., Krusmark, M. A., Jastrembski, T., Hansen, D. A., Honn, K. A., & Gunzelmann, G. (2022). Enhancing learning and retention through the distribution of practice repetitions across multiple sessions. *Mem. Cognit.* ***https://doi.org/10.3758/s13421-022-01361-8***

Weaver, J. R. (1976). The relative effects of massed versus distributed practice upon the learning and retention of eighth grade mathematics. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, *7624394*(5-A), 2698. ***http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc2&NEWS=N&AN=1978-10974-001***

Wechsler, D. (1997). Wechsler memory scale third edition (WMS-III). *San Antonio, TX: The Psychological Corporation*.

Whiffen, J. W., & Karpicke, J. D. (2017). The role of episodic context in retrieval practice effects. *J. Exp. Psychol. Learn. Mem. Cogn.*, *43*(7), 1036–1046. ***https://doi.org/10.1037/xlm0000379***

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing*. Academic press.

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychol. Bull.*, *147*(4), 399–435. ***https://doi.org/10.1037/bul0000309***

Yazdani, M. A., & Zebrowski, E. (2006). Spaced reinforcement: An effective approach to enhance the achievement in plane geometry. *Journal of Mathematical Sciences and Mathematics Education*, *1*(3), 37–43.

Yeo, D. J., & Fazio, L. K. (2019). The optimal learning strategy depends on learning goals and processes: Retrieval practice versus worked examples. *J. Educ. Psychol.*, *111*(1), 73–90. ***https://doi.org/10.1037/edu0000268***