# How does procedural complexity affect the efficacy of spaced retrieval practice?

Ewan Murray1, Aidan J. Horner1, and Silke M. Göbel1,2

1University of York

2University of Oslo

# Author Note

# Abstract

This document is a template.

*Keywords*: keyword1, keyword2, keyword3

# How does procedural complexity affect the efficacy of spaced retrieval practice?

Recent advice on best practice in education has 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. Ideas which could have been taught days, months, or years previously. Unfortunately, students often struggle to retrieve information (Karpicke, 2012). Learners’ failure to retrieve past material requires instructors to allocate additional time to reintroduce it, impeding not only their own learning but, in the case of classroom learning, others. 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 the spacing effect and testing effect, respectively. The spacing effect describes how distributing two or more practice sessions over time, rather than in one massed session, affects retention on a post-test. While the testing effect describes the change in retention when material is tested rather than restudied. In combination, they form spaced retrieval practice (Hopkins et al., 2016). Already, spaced retrieval practice is in regular use by some students in the UK. Multiple online programs such as Arc Education (n.d.) and HegartyMaths (n.d.-a) claim to harness the spacing and testing effect to improve retention for pupils. However, there is limited research into the efficacy of these programmes (Gidaropoulos, 2021) with no peer-reviewed studies investigating their use of spacing. With sites such as HegartyMaths having around 5000 daily users (HegartyMaths, n.d.-b), the majority being students in class or at home, a small positive effect in retention could have a large impact. Furthermore, there is increasing push to bring these effects into classroom use. A recent report by the British government suggests frequent, low-stakes testing as an essential strategy (Ofsted, 2021). Understanding when, how and what factors moderate the use of spaced retrieval, before widespread use, is essential. The purpose of this literature review is, first, to review the existing evidence for spaced retrieval practice for mathematics and, second, to explore what factors moderate the effectiveness of spaced retrieval practice in mathematics learning. The evidence for the spacing and testing effect generally 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 of the effectiveness of the combination of spacing and testing (spaced retrieval practice) in mathematics will be reviewed along with proposed moderators of its effectiveness. Subsequently, there will be a discussion of complexity, why it is important and how it is defined in relation to mathematical content. Following these definitions, the potential effects complexity may have on the efficacy of spaced retrieval practice, and how it may interact with the different accounts of the underlying mechanisms of spacing and testing effects will be explored. Finally, there will be a short discussion of possible hypotheses and experiments. # Spacing effect and distributed practice This section will briefly outline the history of the spacing effect, and how the effect has been studied across domains, before returning to mathematics later. 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 over time, 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 undertaken by Hattie (2008) into many educational interventions found spaced versus massed practice to be one of the most effective interventions. They found an effect of d = 0.71, using two meta-analyses containing 63 studies and 5,028 participants. As an example of what a Cohen’s d = 0.7 effect suggests, if a study which claimed to raise IQ (where the mean is 100 and standard deviation is 15) then a 0.7 effect size would mean a 10.5 point increase in IQ. The history of the spacing effect probably began in parallel with the advent of experimental psychology. Ebbinghaus (1885/1913), for example, 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); 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). Most prior experiments into the spacing effect have investigated learning verbal material (Delaney et. al., 2010, provides a critical review), but there has been growing research across other domains, including mathematics, which will be discussed in a later section. Figure 1 Forgetting curve showing the change in retrieval strength as a function of time A picture containing line, text, plot, diagram

Description automatically generated 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 faded red dashed lines show how the memory would have continued to decay without the retrieval events.

Spacing effect experiments are normally run within or across-session (Küpper-Tetzel, 2014). Where within-session experiments look at the temporal spacing of stimuli within a single session and across-session experiments look at the spacing of practice across multiple sessions. Within mathematics there is very little work done on within-session spacing (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 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 randomization, the only difference between the two conditions is the temporal spacing. Figure 2 Simple spacing paradigm A picture containing text, screenshot, diagram, rectangle

Description automatically generated Note. 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. One large-scale online study (n = 1,350) taught participants 32 obscure facts (Cepeda. et. al., 2008). They varied the inter-session interval from 7 to 105 days and the retrieval interval from 7 to 350 days to optimally schedule the learning of facts. They 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 (Cepeda. et. al., 2008). This relationship between the inter-session interval and the retrieval interval has been named the Glenberg surface (Delaney et al., 2010). It refers specifically to the non-monotonic relationship between the length of the inter-session interval and of the 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 depends on for 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 or retrieval after the next longer gap (Rea & Modigliani, 1985). And 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 the two schedules offer equal benefit, then expanding schedules are more time efficient. As expanding schedules produce the same gain in retention over the same period of time with fewer practice sessions. Alongside increased retention, spacing also improves students’ and teachers ability to accurately gauge students’ learning. When year 7 students were asked to predict their scores on a post-test, after completing either a massed or spaced practice routine, spaced practice increased both the scores of the pupils and the accuracy of their predictions (Emeny et al., 2021). While those on the massed schedule were overconfident. Emeny et al. (2021) suggest the overconfidence may have arisen either because massed practice led to greater fluency within the session, however this performance did not lead to greater long term learning. ## 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; Küpper-Tetzel, 2014; Walsh et al., 2018). There are many theories that aim to explain the spacing effect, however, currently no one theory or simple 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 (Thios & D’Agostino, 1976) suggests that the memory of a previously studied item is strengthened by the retrieval of the original learning event. The degree to which study-phase retrieval improves retention is dependent on the difficulty of recall (Küpper-Tetzel, 2014). The harder the item is to recall, while still being successful, the better for future recall. This is in line with the idea of desirable difficulties, which suggested that there are techniques (e.g., spacing, interleaving, retrieval practice) that make learning harder, to the detriment of performance, but leads to greater long term retention (Bjork et al., 2011). Evidence for this phenomena 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 in a spacing experiment, the probability of successful retrieval is too low and therefore in the study-phase retrieval account, retention for that item is worse. 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 how far apart two presentations of the same word, but not two differing words (Hintzman & Block, 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 one list. Another theory to account for 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 (i.e., Dempster, 1987). A particularly notable early failure to empirically test this theory was performed by (Ross & Landauer, 1978). In this experiment they had 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 datasets 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 deficient processing account of the spacing effect suggests that the phenomenon arises due to learners not processing the material in sufficient depth under the massed condition (Delaney et al., 2010). 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 little evidence to support a claim that deficient processing is the sole mechanism for the spacing effect. Numerous models based on the above theories have been created. One experiment compared three computational models, each trained on 14 previous spacing experiment datasets 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, 1974). 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), instead 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.

# Testing effect and retrieval practice

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 (see figure 3). 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 a effect size (g = 0.499) (Yang et al., 2021). To illustrate what an 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. Figure 3 Simple testing paradigm A picture containing text, screenshot, font, line

Description automatically generated Note. A simple testing paradigm only changes the type of task used after the initial learning session. A retrieval task is compared to a non-retrieval (review) task.

# Theories of the testing effect

There are many theories of the spacing 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 successful retrieval, rather than restudying it, causes 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 tested it against the datasets 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 either strongly associated, they give the example of 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 (Morris et al., 1977, Blaxton, 1989). 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. Later on a free recall task, participants who made the temporal judgements on the words performed significantly better. This supports the episodic context account. # 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 (see below for details). Furthermore, there is evidence that spaced retrieval can greatly increase learning. One recent meta-analysis found a large overall effect ( 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 about 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. ## Retrieval in mathematics There is limited evidence specific to the domain of mathematics for the testing effect with no spacing. An exception is a well-designed study by Fazio (2018) comparing retrieval practice with worked example problem pairs (Fazio, 2018). 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. Some spacing experiments found that it only increased the performance of medium performers (Nazari & Ebersbach, 2019). One reason, not discussed by the authors, may be that for many experiments involving mathematics material it is difficult to know whether pupils actually had to retrieve the material or whether they could potentially have used the resources provided to mitigate the need to retrieve in the “practice” conditions. Here are two examples where authors are specific about what participants had access to: “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). In both cases it is plausible that participants who were struggling looked back at their previous work, not having to retrieve the information and effectively restudying the problem, possibly negating the spacing effect. This may make (spaced) retrieval practice in mathematics most useful for learners who grasped the topic enough to at least attempt to retrieve the procedure/ concept before looking back at previous work and examples. This provides increased benefit as, even if they are unable to successfully retrieve the procedure, prior research has found that even incorrect retrievals can improve future retrieval (Potts et al., 2019). Indeed, exploratory analyses of an experiment that failed to find an overall spacing effect have found that distributed practice may be more effective for students of medium performance level (Nazari & Ebersbach, 2019). They suggest their results were due to low performers not managing to grasp the idea while the high-performance group was immediately at ceiling and therefore the practice condition had no visible effect. The need to ensure adequate practice for low performers and then force retrieval within retrieval and spaced retrieval experiments is important, as it may account for the experiments where no effect was found for those groups (Barzagar Nazari & Ebersbach, 2019).

## Typical paradigms in mathematics spaced (retrieval) experiments

The literature which underpins this project falls broadly into three possible 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 (for illustrative examples, see Roher & Taylor, 2006, 2007), 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 (Lyle, 2020; (Bego et al., 2017; Crissinger, 2015)). While lacking ecological validity, the isolated inter-session interval/retrieval interval experiments (see figure 2) provide a purer measurement of the spacing effect than within course experiments as rearranging 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 et al., 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 & Taylor, 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 particular 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 very 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). ## Moderators of efficacy 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, 2018). 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. This definition is underspecified and does not provide a concrete measure of complexity but does suggest complexity may be an important moderator. The next section discusses complexity in mathematics and ways to measure complexity. # Complexity in Mathematics Complexity has been chosen as the first moderator of the spacing effect to investigate for three reasons. Firstly, as Donovan and Radosevich (1999) found that it was a significant moderator of the spacing effect in their meta-analysis. Secondly, the literature review undertaken so far has not found any studies that systematically investigate complexity. 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. ## Defining complexity Complexity is often defined procedurally or conceptually (Crooks & Alibali, 2014). Previous reviews of the literature surrounding mathematical complexity of 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 defined conceptual knowledge (Crooks & Alibali, 2014). Due to the lack of consensus on defining and operationalising conceptual complexity this project will focus on procedural complexity. While higher-level conceptual knowledge is important, basic procedural facts are still vital. For example, being able to quickly retrieve a procedure or basic fact is very 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 by definition, 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 very 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 & Taylor, 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. ## Complexity and spacing Several theories of the spacing effect may be affected by 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 was 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 suboptimal 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. An experiment that implements this 2 (low or high complexity) x 2 (spaced or massed practice) x 2 (shorter or longer inter-session interval) design could help make the spacing effect more effective for a wider range of material by helping to shape learning schedules and provide evidence for or against the study-phase retrieval hypothesis. 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. Similarly, 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. 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. 5. 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 moderator due to past meta-analyses (Donovan & Radosevich, 1999), the lack of any studies that systematically investigate complexity and the ease in which it could be used to improve current spaced repetition algorithms. Therefore, initially this project will look at the following research questions and investigate these hypotheses: Research questions: 1. Does spaced retrieval practice work for mathematics learning? 2. Does the procedural complexity of the material affect the efficacy of spaced retrieval practice? 3. How does the procedural complexity of the material affect the efficacy of spaced retrieval practice? Hypotheses: 1. Spaced retrieval practice increases retention of mathematics learning. 2. Spacing schedules that are best for lower complexity material will not benefit higher complexity material 3. Decreasing the inter-session interval will make spacing schedules more effective for higher complexity material Planned studies: \* Meta-analysis: \* A meta-analytic review of the effectiveness of spacing and retrieval for the retention of mathematics. \* Will aim to address RQ 1 and look for variables that moderate spaced retrieval. \* Experiment 1 \* A two (spaced/ massed practice) between by two (low/ high complexity) within subjects design aimed to answer RQ 2. \* Experiment 2 \* A two (spaced/ massed practice) between by two (shorter longer interval) between by two (low/ high complexity) within subjects design aimed to answer RQ 3. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Appendix

# Title for Appendix